NBER WORKING PAPER SERIES

SYSTEMIC DISCRIMINATION: THEORY AND MEASUREMENT

J. Aislinn Bohren
Peter Hull
Alex Imas

Working Paper 29820
http://www.nber.org/papers/w29820

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2022, Revised December 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, Dami an 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). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Systemic Discrimination: Theory and Measurement
J. Aislinn Bohren, Peter Hull, and Alex Imas
NBER Working Paper No. 29820
March 2022, Revised December 2024
JEL No. D63, D83, J16, J71

ABSTRACT

Economists often measure discrimination as disparities arising from the direct effects of group identity. We develop new tools to model and measure systemic discrimination, which instead captures how discrimination in other decisions indirectly contributes to disparities. We propose an experimental design, the Iterated Audit, to identify systemic discrimination. We then illustrate these new tools in two field experiments. The first experiment shows how racial discrimination accumulates across multiple rounds of hiring through the interaction of two forces: greater discrimination against inexperienced workers—which affects the opportunity to obtain experience—and high subsequent returns to experience. The second experiment shows how gender-based differences in the language of recommendation letters can translate into systemic gender discrimination in STEM hiring. We discuss how our findings qualify previous results on direct discrimination and outline how our tools can be used to target policy interventions.

J. Aislinn Bohren
Department of Economics
The Ronald O. Perelman Center for
Political Science and Economics
University of Pennsylvania
133 South 36th Street
Philadelphia, PA 19104
abohren@gmail.com

Alex Imas
Booth School of Business
University of Chicago
5897 S. Woodlawn Avenue
Chicago, IL 60637
and NBER
alex.imas@chicagobooth.edu

Peter Hull
Department of Economics
Box B, Brown University
Providence RI 02912
and NBER
peter_hull@brown.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, as well as 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; De Plevitz 2007; Powell 2008; Small and Pager 2020). Decades of research in labor economics similarly notes how "pre-market" discrimination in education and housing might affect the employment opportunities of minorities even in the absence of direct discrimination (Cain 1986; Neal and Johnson 1996; Bertrand and Mullainathan 2004; Blank 2005).[1] More recently, computer scientists have shown how discrimination in algorithmic decisions can arise indirectly from biased data collection and training even when the algorithm is "blinded" to protected characteristics (Angwin et al. 2016; Rambachan and Roth 2020). But despite these insights, when compared with the robust toolkit for modeling and measuring direct discrimination, economists have more limited theoretical and empirical tools 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 empirical tools for bringing this framework to 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] To the

[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. Our framework provides a decomposition of this unconditional difference into different forms of systemic discrimination as well as potentially non-discriminatory disparities, and allows for measurement of the different systemic channels.

[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.

labor economics literature on pre-market discrimination and the computer science literature on algorithmic fairness, we contribute a general approach for formally modeling systemic discrimination in a wide range of settings. 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 different systemic factors on observed disparities and to help inform potential policy responses.

We start by developing a general framework for examining direct and systemic discrimination, including how the former can contribute to the latter through interactions across time and 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 relevant attributes (e.g., work experience) at other decision nodes. Within this system, disparities at a given 'focal' node arise from three distinct channels: direct discrimination at that node (e.g., preferring to hire men over women with identical experience), systemic discrimination arising from the impact of direct discrimination at other nodes (e.g., direct discrimination in past hiring leading to differential experience for women vs. men, and hence further hiring disparities), and differences in characteristics that do not arise from discrimination (e.g., gendered differences in innate physical capacity for work). Direct discrimination at other nodes contributes to systemic discrimination at the focal node when it generates differences in the signaling technology (e.g., experience signals productivity) or payoff-relevant characteristics (e.g., experience increases productivity itself), which lead to treatment disparities at the focal node (e.g., experienced workers are more likely to be hired).

This framework highlights a key analytic choice that researchers interested in quantifying systemic discrimination must make: which systemic forces to analyze. Researchers may wish to hone in on a subset of nodes in order to isolate how decisions within this subset contribute to disparities at the focal node—perhaps due to the availability of detailed decision data for these nodes, policy-makers' capacity to intervene at these nodes, or simply because these are the systemic forces of interest. This amounts to a choice of decision nodes to include in the analysis, which we refer to as the *subsystem*, and a measure to group comparable workers at entry into this subsystem, which we refer to as the *reference qualification*.

For example, a researcher interested in studying systemic discrimination in hiring (the focal node) stemming from direct discrimination by reference letter writers would include both the hiring and reference letter nodes in the subsystem. This would allow her to measure how direct discrimination in reference letters translates to disparities in entry-level hiring. Another researcher interested in the impact of discrimination in both internship opportunities and letter writing on hiring would study a 3-node subsystem that also includes the internship node. Finally, a third researcher interested in studying all informational sources of systemic

discrimination would include all nodes that generate disparities in the information available at the hiring node for workers with the same hiring-node productivity. These approaches contrast with a researcher only interested in direct discrimination in hiring, who would include only this node in the subsystem. What we refer to as *total* discrimination aggregates disparities at the focal node stemming from systemic discrimination within the subsystem and any focal-node direct discrimination.[3] Thus, by choosing different subsystems, the framework breaks down the complex system that generates disparities into interpretable and measurable components and gives a unified structure for different notions of discrimination.[4]

We next show how this framework can be brought to data with the Iterated Audit (IA) approach. An IA analysis involves iterating decisions across multiple nodes. It has two components: (i) a treatment component capturing direct discrimination at 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. Consider, for example, a two-node subsystem 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 and entry-level hiring, and (ii) how internship experience impacts entry-level hiring for individuals with otherwise similar resumes. The latter interaction component shapes how direct discrimination in internship hiring drives systemic discrimination in entry-level hiring.

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* IA 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* IA approach directly measures systemic discrimination by simulating the distribution of focal-node signals, as impacted by group membership and the actions at other 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 application of

[3]Any remaining disparity at the focal node may either stem from systemic discrimination arising from nodes outside of the subsystem or non-discriminatory factors.

[4]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 the author notes, 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, as well as separates the first component into direct and systemic components.

the constructive approach 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 undo any perceived direct discrimination at other nodes.

We illustrate these new tools in two field experiments. The first experiment 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.[5] Here, we considered 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 2004; Kline et al. 2022) 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. (2022) document as having the highest levels of direct discrimination in callback rates, and similarly randomize applicants' names to signal their race.[6] 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.[7]

Estimating the interaction component, and hence systemic discrimination, requires assessing (i) 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

[5]The impact of entry-level disparities on later job-market outcomes is highlighted in the 2017 statement by the Association of Women Surgeons: "The disparities women face in compensation at entry level positions lead to a persistent trend of unequal pay for equal work throughout the course of their careers."

[6]Kline et al. (2022) ran a correspondence study across a large set of US firms. The study used fictitious resumes that were identical except for the applicant's name, which was either distinctly Black or distinctly White. Importantly, each resume had at least some level of experience. The authors observed a significant 5.3 percentage-point callback gap between Black and White resumes sent to automotive firms.

[7]This gap is higher than previous estimates of callback effects which, to the best of our knowledge, used resumes with prior experience (Bertrand and Mullainathan 2004; Nunley et al. 2015; Deming et al. 2016).

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.

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 about 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 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 or 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 lab-in-the-field experiment demonstrates the value of the experimental IA method in settings with high-dimensional or complex signals. In this study, we measured how direct gender discrimination in the language of recommendation letters can generate systemic discrimination in hiring. As before, we considered a system with two nodes. Applicants receive 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 et al. 2007), which are replicated in automated recommendation letters from large language models (LLMs) (Wan et al. 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 decisions. 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 across 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) followed a standard correspondence study design, with set A assigned distinctively male names, set B assigned distinctively female names, and the recommendation letters in both sets corresponded to those originally 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 et al. 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 found that essentially all of the gender discrimination in hiring rates and wages was 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 findings are in line with recent work on gender discrimination in labor markets which suggests a limited role of direct discrimination in explaining observed gender disparities. For example, a recent meta-analysis of correspondence studies in Schaerer et al. (2023) found little discrimination in aggregate when non-group attributes are held fixed (though there was substantial heterogeneity across employers/sectors, with some favoring women and others favoring men; see also Kline et al. 2022 and Ceci and Williams 2011). Our results suggest that systemic discrimination arising from direct discrimination in how male and female candidates are described to employers can potentially explain some of the total disparities observed in the labor market that conventional direct discrimination measures miss.

This paper builds on several related literatures in economics. The notion of pre-market discrimination in labor economics is 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.,

Gynter 2003; Feagin 2013) as well as more modern notions of indirect economic discrimination (e.g., Hurst et al. 2024). Connecting these different literatures yields a general approach to measurement and a unified framework for considering appropriate policy responses. More recently, McMillon (2024) provides a taxonomy to delineate different ways an initial instance of discrimination can propagate within a discriminatory system. He also suggests that such systems may be exploited to amplify the effects of equity-focused interventions.

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). Indeed, the question of what to control for in a regression of outcomes on group membership is similar to the choice of subsystem and reference qualification. This choice leads naturally to the specification of "good" and "bad" controls in our framework. Given a set of "good" controls, our framework also decomposes the resulting conditional disparity into direct and systemic components, and directly links the systemic component to direct discrimination at particular other decision nodes. Discrimination measures with more and fewer controls can thus be reconciled as identifying different forms of systemic and total discrimination.

Finally, our experimental findings add to the growing literature estimating the impact of previous direct discrimination on subsequent disparities (Cook 2014; Williams et al. 2021; Derenoncourt et al. 2024; Eli et al. 2023; 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 (Zivin and Singer 2022; Lodermeier 2023; Gawai and Foltz 2023; Buchmann et al. 2023; Abramitzky et al. 2023; Althoff and Reichardt 2024; Baron et al. 2024). More broadly, we join a literature modeling and estimating the indirect impact of discrimination on important economic outcomes—including Darity (2005), Bohren et al. (2019), Arnold et al. (2022), Bohren et al. (2023), and Hurst et al. (2024).

We organize the remainder of this paper as follows. Section 2 presents a simple example motivating a systems-based approach to discrimination. Section 3 develops our theoretical framework for studying direct and systemic discrimination. Section 4 discusses measurement, including the IA design. Sections 5 and 6 present the empirical applications. Section 7 concludes. Appendix A reviews connections to related literatures, Appendix B presents an additional application, and Appendix C presents further details of the empirical studies.

2 Motivating Example

We start with a simple example that illustrates key features of the framework. Consider a population of computer programmers applying to a job at a software company to code a new feature. Let $Y_i^* \in \{0,1\}$ indicate whether, if hired, programmer i would be able to successfully code the feature and let $A_i^* \in \{0,1\}$ indicate whether programmer i is hired.

To hire programmers, the company solicits a resume R_i and a performance profile P_i for each applicant, where P_i is a rating of performance on coding assignments the applicant completed on an online coding platform such as GitHub. Let C_i denote the code i submitted to the platform, which forms the basis of his or her rating P_i.[8] The company's hiring manager observes this resume and performance profile for each programmer i, denoted by $S_i^* = \{R_i, P_i\}$, as well as the programmer's gender G_i.

A large literature documents gender-based disparities in hiring in STEM fields (e.g., Davison and Burke 2000; Roth et al. 2012). Suppose three economists are interested in studying the role of discrimination in generating such disparities. Economist 1 follows standard practices in the field with a carefully designed correspondence study. Specifically, she sends the company a set of fictitious resumes and performance profiles $S_i^* = \{R_i, P_i\}$, along with a signal of each fictitious applicant's gender G_i (e.g., a distinctively male or female first name), and elicits a hiring decision A_i^*. By randomizing G_i, she is able to estimate the causal effect of gender on hiring decisions conditional on the resume and profile S_i^*. She finds that male applicants are slightly less likely to be hired than female applicants with the same resume and profile. She concludes that there is some discrimination against male applicants.

Economist 2 is interested in the same question, but ends up running a slightly different correspondence study. Rather than generating fictitious performance profiles, she generates code for each fictitious applicant and submits it to the *online platform* (along with the salient signal of the applicant's gender). The performance profile is generated by evaluators on the platform based on this code. The economist then submits fictitious resumes and these performance profiles to the company. Thus, while applicants have comparable resumes and code, their performance profiles may differ based on how they are evaluated on the platform. Strikingly, she reaches a different conclusion than Economist 1: male applicants are slightly *more* likely to be hired than female applicants with the same resume and submitted code. She concludes there is some discrimination against female applicants in this setting. In unpacking this result, she finds that evaluators on the platform tend to be less willing to accept the same code from female programmers than male programmers, which generates gender-based disparities in the resulting profiles.[9]

Finally, Economist 3 examines the same question by running a different type of study. She recruits a real set of male and female programmers with similar resumes and performance

[8] Programmers' contributions to Open Source Software (OSS) projects often play a role in labor market hiring decisions. OSS projects are typically hosted by project managers on platforms such as GitHub, where project managers post 'pull requests' which allow other users to contribute code for specific aspects of the project. The project manager evaluates this code and decides to accept or reject it. A user's history of accepted pull requests is part of their GitHub profile, which is often considered in hiring decisions. Prior work has documented gender-based disparities in the acceptance of pull requests on GitHub (Terrell et al. 2017), as well as in performance profiles on other coding platforms, e.g., Stack Overflow (May et al. 2019).

[9] See Terrell et al. (2017) for evidence of such disparities.

profiles as the fictitious applicants in the first two studies. She then hires them to code a new feature similar to the one desired by the company. In this way, Economist 3 is able to measure the coding ability Y_i^* of each programmer. Submitting these programmers' resumes and performance profiles to the company, she computes gender disparities in elicited hiring decisions among applicants with the same true coding ability—without conditioning on the resume, code submitted to the platform, or performance profile. Curiously, she reaches a different conclusion than both Economist 1 and 2: male computer programmers are *much* more likely to be hired and receive a positive evaluation than female computer programmers with the same underlying ability. In unpacking this result, Economist 3 finds that a key driver is differential prior experience: among programmers with the same coding ability, female programmers were hired for far fewer coding jobs.

At first blush, this simple example presents a puzzle: which of the three researchers are correct on the nature and extent of discrimination in STEM hiring? Economist 1 follows the norm in audit and correspondence studies by conditioning on the information S_i^* available to the decision-maker of interest (i.e., the software company). Economist 2 finds that some of this signal is biased by a different evaluation (i.e., the performance profile); her measure of discrimination takes this bias into account. By conditioning on an "objective" measure of coding ability Y_i^*, Economist 3 uncovers a further bias in the applicant's opportunity to gain experience. How can the *systemic* biases that Economists 2 and 3 find be studied alongside the *direct* discrimination Economist 1 finds?

In Section 3 we develop a general framework that reconciles these analyses. The framework formalizes how the study of discrimination beyond a single decision requires a researcher to select a set of other decisions to include in the analysis—what we call the subsystem—which is the key factor that differed across the three economists here. In any given setting, by selecting different subsystems, 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 generally requires new empirical tools. While Economist 1 identified direct discrimination with a standard correspondence study, isolating the effects of the systemic biases found by Economists 2 and 3 is more challenging. We propose an new Iterated Audit design to identify these effects.

Second, this example highlights how it may be difficult to address such systemic forms of discrimination with standard individual-level interventions. The hiring managers at the software company in this example seem at least partly aware of the systemic bias in performance profiles, given their "reverse" discrimination in evaluating female versus male applicants. Yet they do not fully offset this bias, either because of imperfect awareness, psychological frictions, or their own 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 system-wide policy responses and study how they may impact other interconnected decisions.

3 A Framework for Systemic Discrimination

We now develop a theoretical framework to formalize a notion of systemic discrimination. Section 3.1 introduces the setting and defines three forms of discrimination: direct, total, and systemic. Section 3.2 relates these via a decomposition of total discrimination into direct and systemic components. Section 3.3 discusses key features of the framework: how it relates to notions of systemic discrimination in other fields, how it nests existing notions of discrimination in the economics literature through the choice of a subsystem, the sources of systemic discrimination, and the relationship between systemic and statistical discrimination.

3.1 Framework and Definitions

We develop the framework in a labor market context, in which we consider discrimination among workers being evaluated for a task n^*. Worker i has ex-ante unobservable *productivity* $Y_i^* \in \mathcal{Y}^*$ on the task (e.g., error rate, number of units produced), where $\mathcal{Y}^*$ is the set of possible productivity levels. A manager observes the worker's *group membership* $G_i \in \{b, w\}$ and a vector of other attributes $S_i^* \in \mathcal{S}^*$ (e.g., educational background, recommendation letters) which we refer to as the *signal*, where $\mathcal{S}^*$ is the set of possible attribute vectors. A worker's attributes and group identity provide information about her productivity. 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.[10] 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 action rule $A^* : \mathcal{S}^* \times \{b, w\} \to \mathcal{A}^*$, which determines how the observed signal and group identity maps into an action choice.[11] Given G_i and S_i^*, the manager selects action $A_i^* = A^*(G_i, S_i^*)$.

We embed this labor market decision in a broader economy—a *system*—to capture the idea that a worker's productivity and signal for task n^* may be affected by actions in other domains (e.g., financial, criminal, or past employment) or time periods. The system consists of a set of nodes $\mathcal{N} \equiv \{1, ..., N\} \cup \{n^*\}$, where we refer to n^* as the *focal node*. Each non-focal node $n = 1, ..., N$ corresponds to a task similar in structure to n^*. Specifically, at each n, a

[10] We assume the action is a real number to focus on expected action as a measure of discrimination and we assume the set of groups is binary to simplify notation; the analysis easily extends to more general spaces.

[11] The framework can accommodate a random action rule that induces a distribution over $\mathcal{A}^*$. For simplicity, we focus on a deterministic action rule, as well as abstract from interactions across workers and other realistic features of labor markets.

worker is evaluated for a task for which he has productivity $Y_i^n \in \mathcal{Y}^n$ (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 G_i and a signal $S_i^n \in \mathcal{S}^n$, then selects action $A_i^n \in \mathcal{A}^n$ (e.g., loan terms, criminal charges, performance evaluation). As above, the sets $\mathcal{Y}^n$, $\mathcal{S}^n$, and $\mathcal{A}^n$ give the possible productivity levels, attribute vectors, and actions, respectively, and $A^n(G_i, S_i^n)$ denotes the evaluator's action rule.[12] At the end of this subsection we discuss two additional types of decisions this system can include: *(i)* nodes where the decision is impacted by an anticipated action at a future node, and *(ii)* nodes where the decision is the choice of signaling technology for another node.

Actions at one node can impact the productivity and signal at others. For example, the signal S_i^* at the focal node n^* may include a performance evaluation A_i^n from another node n. Similarly, focal-node productivity Y_i^* may be a function of productivity Y_i^n and on-the-job training A_i^n at node n. For simplicity, we maintain that group membership G_i remains the same across all nodes.

We first define *direct discrimination*, which captures group-based differences in focal-node 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 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 $A^n(w, s) \neq A^n(b, s)$.*[13]

Direct discrimination arises from the worker's group identity itself. It can arise from the manager's preferences, beliefs about productivity, or beliefs about the signal distribution being dependent on group identity (e.g., taste-based or statistical discrimination). Because this definition conditions on the signal at the node where discrimination is being measured, it does not account for disparities that stem from decisions made at other nodes.[14]

We next aim to define a notion of discrimination that incorporates how discrimination at other nodes contributes to disparities at the focal node. A researcher interested in identifying and quantifying such broader inequities must make a key analytic choice: namely, which systemic forces to study. The researcher may wish to hone in on a subset of nodes in order to isolate how actions within this subset contribute to disparities at the focal node—perhaps due to the availability of detailed decision data for these nodes, policy-makers' capacity to

[12] We don't place any restrictions on the sets $\mathcal{Y}^*$, $\mathcal{S}^*$, $\mathcal{Y}^n$, $\mathcal{S}^n$, and $\mathcal{A}^n$ aside from that it is possible to endow each set with a σ-algebra so that probability measures over each set are well-defined.

[13] 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.

[14] We abstract from some important conceptual issues with considering group membership G_i as separately manipulable in the action rule from S_i^*. Such issues can be especially salient when G_i is meant to capture race; see, e.g., Fryer and Levitt (2004), Sen and Wasow (2016), Gaddis (2017), and Kohler-Hausmann (2019). Notably, the constructivist framework of Rose (2023) can be incorporated to address this issue given a salient and manipulable signal of group membership (e.g., distinctively racial names).

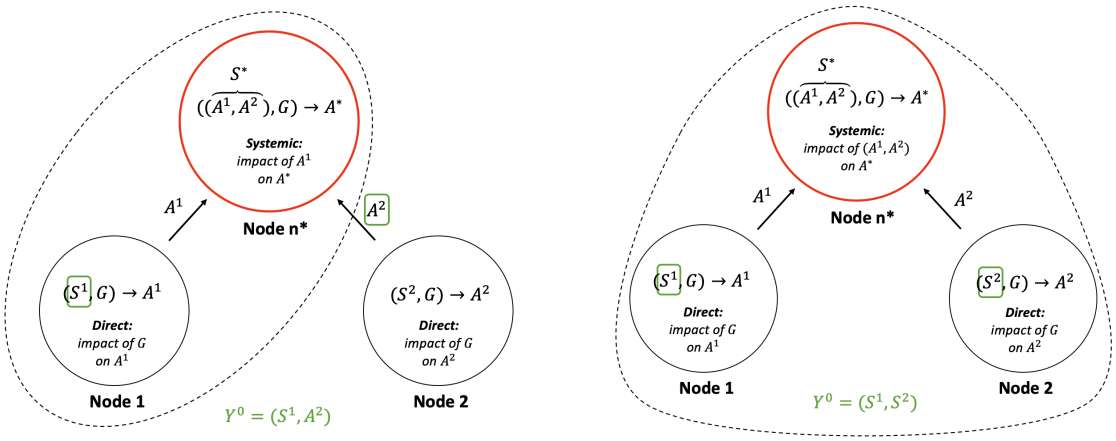
intervene at these nodes, or simply because these are the systemic forces of interest to the researcher. This amounts to choosing a subset of other nodes $\mathcal{N}^0 \subset \mathcal{N}$ to include in the analysis, which we refer to as the *subsystem*, and a measure to group comparable workers at entry into this subsystem, which we refer to as the *reference qualification*. Given a chosen reference qualification, let Y_i^0 denote worker i's qualification level and $\mathcal{Y}^0$ denote the set of possible qualification levels.

Figure 1 illustrates two possible choices of subsystem in a three-node system $\mathcal{N} = \{1, 2, n^*\}$. As in Section 2, suppose the focal node n^* corresponds to a hiring decision for a software company and node 1 corresponds to a platform performance rating (e.g., a GitHub profile). The platform signal S_i^1 corresponds to the programmer's submitted code and the platform action A_i^1 corresponds to her rating. Suppose node 2 corresponds to a software development internship, where the programmer interviewed for the position (signal S_i^2) and either obtained experience or not, $A_i^2 \in \{0, 1\}$. The hiring node signal then consists of the platform rating and the presence or absence of internship experience: $S_i^* = (A_i^1, A_i^2)$. A researcher interested in studying how bias in the platform rating (node 1) contributes to hiring disparities (node n^*) would include these two nodes in the subsystem: $\mathcal{N}^0 = \{1, n^*\}$, as illustrated in Panel (a). Setting the reference qualification as the platform node signal (submitted code) and internship node action (internship experience, since this node is not part of the subsystem) groups comparable workers at entry to this subsystem: $Y_i^0 = (S_i^1, A_i^2)$. This is the implicit choice of Economist 2 in Section 2. If the researcher wants to also study how direct discrimination in internship experience impacts hiring disparities, she would also include node 2 in the subsystem: $\mathcal{N}^0 = \{1, 2, n^*\}$, as illustrated in Panel (b). Now, setting the reference qualification as the platform node signal and internship node *signal*, $Y_i^0 = (S_i^1, S_i^2)$, groups comparable workers at entry to this subsystem. We further discuss the choice of subsystem and qualification in Section 3.3.[15]

Fixing a subsystem and reference qualification, we define a measure of *total discrimination* that captures group-based differences in actions at the focal node for workers who enter the subsystem with comparable qualifications. Worker i enters the subsystem with qualification level Y_i^0 and is evaluated at each node, which potentially impacts his productivity and signal at subsequent nodes. These interactions generate a distribution over how the focal-node productivity Y_i^* and signal S_i^*—and hence action A_i^*—vary by group for workers who start with the same qualification level. Let $\alpha^*(y^0, g) \in \Delta(\mathcal{A}^*)$ denote this action distribution for a group g worker with qualification level y^0, where $\Delta(\mathcal{A}^*)$ denotes the set of distributions over

[15]Note that within a given system, the choice of reference qualification pins down the subsystem. We emphasize both choices as, conceptually, we view it more natural to first choose which systemic forces to focus on (i.e., the subsystem) and then to decide on the measure that ensures comparability at entry to the subsystem (i.e., reference qualification). This also highlights the importance of modeling the system of interactions which give rise to total and systemic discrimination, as defined below, as these structural relationships are crucial for policy evaluation.

FIGURE 1. Two Subsystem Choices in a Three-Node System



(a) Two-node subsystem $\mathcal{N}^0 = \{1, n^*\}$ capturing systemic impact of direct discrimination at node 1

(b) Three-node subsystem $\mathcal{N}^0 = \{1, 2, n^*\}$ capturing systemic impact of direct discrimination at nodes 1 and 2

$\mathcal{A}^*$, and let $\mu^*(y^0, g) \in \mathbb{R}$ denote the corresponding average action.[16] Total discrimination is the difference between the average actions faced by group w and b workers who enter the subsystem with the same qualification level:[17]

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 $y^0 \in \mathcal{Y}^0$.*

By comparing workers with the same qualification level, total discrimination captures focal node disparities that arise due to decisions within the subsystem. Importantly, we do not claim that the residual focal-node disparity is *not* discrimination; this remaining disparity may stem from decisions outside of the subsystem, non-discriminatory forces, or both. Returning to Figure 1, total discrimination for the subsystem in Panel (a) includes the impact of direct discrimination at node 1 on disparities at node n^*, while in Panel (b) it includes the impact of direct discrimination at both nodes 1 and 2. In Section 3.3 we discuss how, through the choice of subsystem and reference qualification, this definition nests common notions of discrimination in the literature.

Total discrimination includes both the impact of direct discrimination at other nodes and at the focal node.[18] We next define a measure of *systemic discrimination* to isolate

[16] Formally, $\alpha^* : \mathcal{Y}^0 \times \{b, w\} \to \Delta(\mathcal{A}^*)$ and $\mu^* : \mathcal{Y}^0 \times \{b, w\} \to \mathbb{R}$ are mappings from the sets of qualification levels and group memberships to a distribution over actions or an expected action, respectively.

[17] Our definitions based on means easily generalize to other distributional features of α^*, such as variances or higher-order moments. We focus on means for notational simplicity and since these yield the most commonly studied discrimination measures in practice.

[18] That total discrimination includes the impact of focal-node direct discrimination can be seen from (1): the dependence of α^* on A^* means that direct discrimination at n^*—as captured by $A^*(b, s) \neq A^*(w, s)$—can

the component of total discrimination that arises indirectly from discrimination at other nodes. In particular, this measure constructs a counterfactual action distribution to compare average actions when both groups face the same focal-node action rule—thereby shutting down the possibility of direct discrimination. Let $\sigma^*(y^0, g) \in \Delta(\mathcal{S}^*)$ denote the node-n^* signal distribution for a group g worker with qualification level y^0. Note that the action distribution $\alpha^*(y^0, g)$ depends on $\sigma^*(y^0, g)$ and the action rule $A^*(g, s)$ for group g:

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

for any measurable set $A \subset \mathcal{A}^*$.[19] Define the counterfactual action distribution $\tilde{\alpha}^*(y^0, g) \in \Delta(\mathcal{A}^*)$ for a group g worker with qualification level y^0 as the action distribution under signal distribution $\sigma^*(y^0, g)$ and action rule $A^*(g', s)$ for group $g' \neq g$:

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

for any measurable set $A \subset \mathcal{A}^*$. Let $\tilde{\mu}^*(y^0, g)$ denote the corresponding average action. Systemic discrimination compares the average action $\mu^*(y^0, w)$ of group w workers to the counterfactual average action $\tilde{\mu}^*(y^0, b)$ of group b workers under the group w action rule—in other words, the average action for group b versus w when both groups face the group w action rule (or vice versa)—holding fixed the qualification level:

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

Varying the choice of subsystem and reference qualification provides measures of systemic discrimination that capture the impact of different sets of systemic forces. While systemic and total discrimination are defined for a particular qualification level, it is also possible to construct an overall measure by averaging across qualification levels.[20]

To illustrate the definitions, consider the analysis of Economist 2 in Section 2. As discussed above, she selects a two-node subsystem with the hiring node (focal node n^*) and the platform node, $\mathcal{N}^0 = \{1, n^*\}$, and she sets the platform signal (i.e., code submitted to the platform) as the reference qualification: $Y_i^0 = S_i^1$. Suppose the platform's performance pro-

translate to different action distributions.

[19] Given a distribution f with support $\mathcal{X}$ parameterized by (y^0, g), we write $f(y^0, g)$ to denote the distribution over $\mathcal{X}$, where $f(y^0, g) \in \Delta(\mathcal{X})$, and $f(X; y^0, g)$ to denote the probability mass $f(y^0, g)$ assigns to measurable set $X \subset \mathcal{X}$, where $f(X; y^0, g) \in [0, 1]$.

[20] The interpretation of this overall measure depends on the qualification distribution used to take the average: using the population qualification distribution measures average discrimination across both groups, while using the qualification distribution for group g measures average discrimination for group g workers.

file is a rating of 0 to 100 (e.g., a code acceptance rate) and that gender bias on the platform leads female programmers to receive systematically lower ratings than male programmers for similar code. Suppose further that the company hires male programmers with a rating above 90 and female programmers with a rating above 80. Thus, there is direct discrimination at both the platform node (from the biased rating, favoring male programmers) and the hiring node (from the differential hiring thresholds, favoring female programmers).

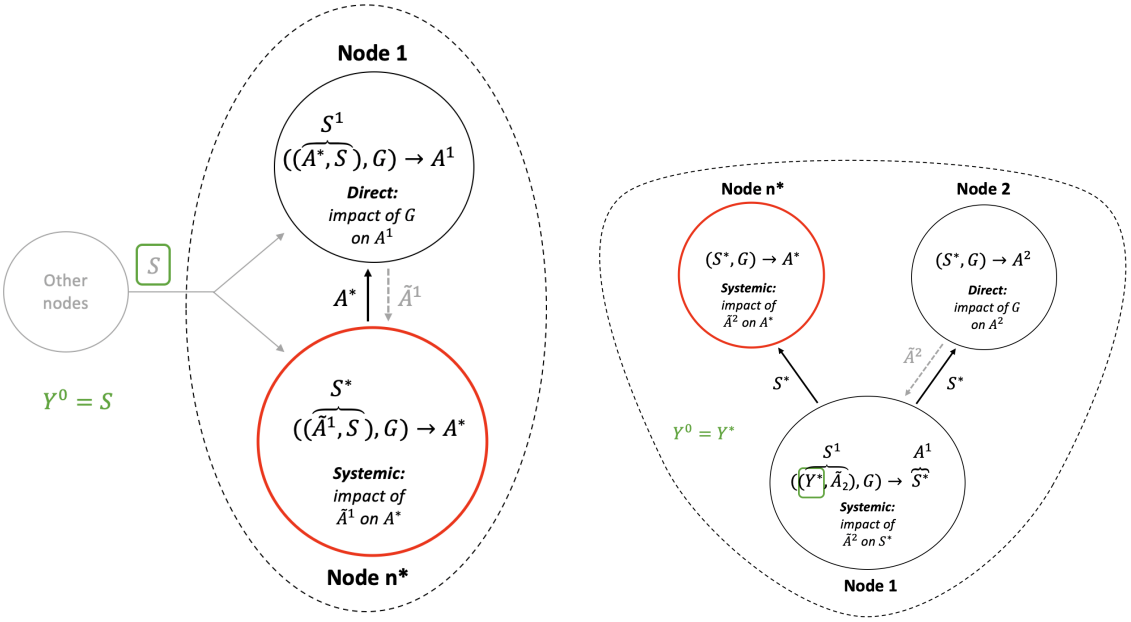
Systemic discrimination captures hiring disparities due to the direct discrimination at the platform node: fixing the hiring threshold as either that used for males (90) or females (80), it compares the probability that females and males who submit similar code have a rating above this threshold. Because of the direct discrimination in ratings, this hiring probability is higher for male programmers than female programmers. Total discrimination combines this with the direct discrimination in hiring: it compares the probability that a male programmer has a rating above 90 to the probability that a female with similar code has a rating above 80. Economist 2 finds that this hiring probability is higher for male programmers than female programmers, because the systematic bias in platform ratings outweighs the 'reverse' direct discrimination in hiring thresholds.

Additional Types of Decision Nodes. Before proceeding, we make two notes on the generality of this framework. First, while the examples so far have focused on systemic discrimination from past actions, our framework can also capture systemic discrimination driven by how anticipated actions at contemporaneous or future nodes impact the focal-node decision. For example, the expected ruling of a jury may impact a defense attorney's decision to accept a plea deal. This can be captured by including the anticipated action $\tilde{A}_i^n$ as part of the focal-node signal S_i^*. This anticipated action may be correct, as in a Nash equilibrium, or incorrect, as in alternative solution concepts (e.g., self-confirming or Berk-Nash equilibrium). Figure 2(a) depicts a subsystem with this type of node. The decision at node 1 is made subsequent to the decision at node n^*, and the anticipated action $\tilde{A}_i^1$ is relevant for A_i^*. The reference qualification (in green) is the component of the node-n^* signal generated outside of the subsystem (i.e., S_i^* excluding $\tilde{A}_i^1$). Systemic discrimination captures the impact of expected direct discrimination at node 1 via $\tilde{A}_i^1$ on A_i^*.[21]

Second, while we have so far focused on settings where systemic discrimination arises from differences in the distribution of a given set of signals (i.e., for a fixed evaluation criteria), the framework can also capture systemic discrimination arising from the choice of the productivity signaling technology itself (i.e., choosing the evaluation criteria). This can be modeled as a decision node 1, where the signal corresponds to focal-node productivity,

[21] For example, Avery et al. (2023) find that the use of artificial intelligence (AI) in recruitment increases the share of women applicants, as women anticipate less bias when assessed by AI instead of humans. Ruebeck (2024) shows that increasing perceived discrimination by revealing that managers mostly promoted White men lowered the retention of female workers and increased their reservation wages.

FIGURE 2. Other Forms of Systemic Discrimination



(a) Two-node subsystem capturing the systemic impact of predicted direct discrimination at future node 1

(b) Three-node subsystem capturing the systemic impact of predicted direct discrimination at node 2 via evaluation design node 1

$S_i^1 = Y_i^*$, the action corresponds to the focal-node signal, $A_i^1 = S_i^*$, and the (random) action rule maps productivity and group to a distribution $\Delta(\mathcal{S}^*)$ over focal-node signals.[22] Figure 2(b) depicts a subsystem with this type of node. At node 1, the worker chooses a signaling technology (e.g., reported traits on a job placement questionnaire); this signal impacts actions at nodes n^* and 2 (e.g., a job hiring decision and a social evaluation). The anticipated action $\tilde{A}_i^2$ impacts the chosen signaling technology at node 1, and hence, the node n^* action A_i^*. The reference qualification is productivity, $Y_i^0 = Y_i^*$, which is the node-1 signal (e.g., actual traits). Systemic discrimination captures the impact of expected direct discrimination at node 2 on the choice of signaling technology, and hence, the focal-node hiring decision A_i^*.[23]

[22] De Plevitz (2007) and Pincus (1996) discuss this design choice as a channel for systemic discrimination, where an institution is first "designed" by a group in power leading to the development of evaluation criteria that are optimized around the characteristics of this group. For example, by not accounting for the family structure and cultural obligations of the Aboriginal community, De Plevitz (2007) discusses how the "Euro-centric model of teaching" creates systemic barriers in schooling for Aboriginal children. Another example is the practice of excluding women from medical trials, which leads to a less informative signal of the efficacy of new treatments for these groups (Bierer et al. 2022).

[23] This example takes inspiration from Bursztyn et al. (2017), where single women report lower desired salaries and willingness to work long hours on a job placement questionnaire (node 1) when the answers were observable to their classmates, as such traits are viewed as undesirable for women (but not men) on the marriage market (node 2). This then led to disparities in job market evaluations (node n^*), where such traits are viewed as desirable for both groups.

3.2 Decomposing Total Discrimination

The three forms of discrimination are linked by decompositions of total discrimination into direct and systemic components. From Definitions 2 and 3, $\Delta^T(y^0)$ measures total discrimination and $\Delta^S(y^0, w)$ or $\Delta^S(y^0, b)$ measure systemic discrimination at qualification level y^0. From Definition 1, $\delta(s) \equiv A^*(w, s) - A^*(b, s)$ measures direct discrimination at n^* following signal realization s. A measure of *average direct discrimination* for workers with the same qualification level is thus given by the expectation of direct discrimination with respect to the signal distribution for this qualification level:

$$\Delta^D(y^0, g) \equiv \int_{\mathcal{S}} \delta(s) d\sigma^*(s; y^0, g) \tag{3}$$

for $g \in \{w, b\}$, where $\sigma^*(y^0, g)$ is the group-specific signal distribution defined above. Similar to how the systemic discrimination measures depend on group via which action rule is fixed, average direct discrimination depends on group via which signal distribution is fixed.

Given these measures, we have the following result:

Proposition 1. *At each qualification level y^0, total discrimination is the sum of average direct discrimination and systemic discrimination:*

$$\underbrace{\Delta^T(y^0)}_{\textit{Total discrimination}} = \underbrace{\Delta^D(y^0, w)}_{\textit{Avg. direct discrimination}} + \underbrace{\Delta^S(y^0, b).}_{\textit{Systemic discrimination}} \tag{4}$$

Equation (4) follows by adding and subtracting $\tilde{\mu}^*(y^0, w)$ to the definition of $\Delta^T(y^0)$ and rearranging terms.[24] It is in the spirit of Kitagawa (1955), Oaxaca (1973), and Blinder (1973), who relate unconditional disparities to a component explained by observable worker characteristics (e.g., education or labor market experience) and an "unexplained" disparity. These classic decompositions can be viewed as a strategy for measuring direct discrimination, while Equation (4) instead leads to strategies for measuring systemic discrimination and relating it to direct discrimination (see Section 4).

As with the classic Kitagawa-Oaxaca-Blinder approach, there are multiple ways to decompose total discrimination into direct and systemic components. We can also express total discrimination as the sum of average direct discrimination with respect to the group b signal distribution and systemic discrimination under the group w action rule:[25]

$$\Delta^T(y^0) = \Delta^D(y^0, b) + \Delta^S(y^0, w). \tag{5}$$

[24] Note that $\mu^*(y^0, w) = \int_{\mathcal{A}^*} a d\alpha^*(a; y^0, w) = \int_{\mathcal{S}^*} A(w, s) d\sigma^*(s; y^0, w)$ and $\tilde{\mu}^*(y^0, w) = \int_{\mathcal{A}^*} a d\tilde{\alpha}^*(a; y^0, w) = \int_{\mathcal{S}^*} A(b, s) d\sigma^*(s; y^0, w)$. Then $\Delta^T(y^0) = \mu^*(y^0, w) - \mu^*(y^0, b) + \tilde{\mu}^*(y^0, w) - \tilde{\mu}^*(y^0, w) = \int_{\mathcal{S}^*} A(w, s) d\sigma^*(s; y^0, w) - \int_{\mathcal{S}^*} A(b, s) d\sigma^*(s; y^0, w) + \Delta^S(y^0, b) = \Delta^D(y^0, w) + \Delta^S(y^0, b)$.

[25] This follows by adding and subtracting $\tilde{\mu}^*(y^0, b) = \int_{\mathcal{S}} A(w, s) \sigma^*(s; y^0, b) ds$ to and from the definition of $\Delta(y^0)$ and rearranging terms.

Averaging Equations (4) and (5) yields a third decomposition:

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

where $\overline{\Delta}^D(y^0) \equiv \frac{1}{2}(\Delta^D(y^0, w) + \Delta^D(y^0, b))$ and $\overline{\Delta}^S(y^0) \equiv \frac{1}{2}(\Delta^S(y^0, w) + \Delta^S(y^0, b))$.

3.3 Discussion

Systemic Discrimination in Other Fields. Our formalization of systemic discrimination aligns broadly with indirect forms of discrimination discussed in the sociology literature: i.e., forms of inequality operating indirectly through characteristics beyond group identity and stemming from discrimination in other parts of a system. As discussed in Small and Pager (2020), part of the challenge of assessing different notions of indirect discrimination in prior work is that different terms—such as "structural," "systemic," and "institutional" discrimination—are used to describe this general phenomenon. Pincus (1996), for example, defines structural discrimination as policies "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).[26] Similarly, Powell (2008) defines systemic discrimination as a "product of reciprocal and mutual interactions within and between institutions," both "within and across domains."[27] Our formalization captures both of these definitions: we measure systemic discrimination as disparities under an action rule $A^*(g, \cdot)$ that is fixed across groups, which arise due to discriminatory decisions at other nodes.[28] Our measure also captures both types of indirect discrimination in Feagin (1977)'s typography: past-in-present discrimination, which arises indirectly from discriminatory actions at past nodes (as in Figure 1) and side-effect discrimination, which arises indirectly from discriminatory actions in other domains.[29] Appendix A reviews other connections to literatures on systemic discrimination.

Choosing a Subsystem and Reference Qualification. The interpretation of the systemic and total discrimination measures is tied to the researcher's chosen subsystem and reference qualification. Through these choices, the measures can bring focus to different systemic forces and study different forms of discrimination.

Three examples illustrate this feature of the framework. First, by including only the

[26] 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 typically being prima facie race-neutral. In our framework, any total discrimination that arises from a race-neutral action rule is systemic, as the direct component is zero by definition.

[27] 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.

[28] Our definition also aligns with some notions of institutional discrimination (Small and Pager 2020).

[29] Notably, while Feagin (1977) referred to both forms of discrimination as indirect institutional discrimination, Feagin (2013) refers to both as forms of systemic discrimination.

focal node in the subsystem ($\mathcal{N}^0 = \{n^*\}$) and choosing the focal-node signal as the reference qualification ($Y_i^0 = S_i^*$), total discrimination is aligned with direct discrimination; it does not account for any systemic forces. This is the choice of Economist 1 in Section 2 and corresponds to model (I) in Cain (1986)'s classification of statistical models for measuring discrimination. Second, choosing focal-node productivity as the reference qualification ($Y_i^0 = Y_i^*$) corresponds to including all nodes that impact how workers signal this productivity at node n^*. This aligns total discrimination with what Arnold et al. (2022) argue captures the legal notion of *disparate impact*.[30] It also relates to some measures of algorithmic unfairness, e.g., Berk et al. (2021).[31] This is the choice of Economist 3 in Section 2, who sets Y_i^0 to be coding ability. Finally, including all nodes in the subsystem ($\mathcal{N}^0 = \mathcal{N}$) and setting the reference qualification to a constant presumes that there are no differences prior to entry in the system and all differences emerge from systemic forces. Total discrimination is then the unconditional disparity between groups and the systemic component accounts for the impact of all actions influencing the focal-node productivity and signal. This choice corresponds to model (II) in the taxonomy of Cain (1986).

Other choices can isolate different systemic forces in the economy. To measure the impact of specific past discriminatory decisions on treatment at the focal node (i.e., past-in-present discrimination) one can include the relevant node(s) in the subsystem and set the reference qualification to be information at entry to the subsystem, as illustrated in Figure 1. This is the choice of Economist 2 in Section 2, who sets Y_i^0 to be the information available to the online platform. Similarly, one can include decision nodes from other domains to study how discriminatory practices in one domain impact treatment in another (i.e., side-effect discrimination). Finally, one can include nodes where decisions are made contemporaneous with or subsequent to the focal node, where anticipated discrimination at these nodes impacts action choices at the focal node, as illustrated in Figure 2.[32]

Thus, through the choice of subsystem and qualification, Definitions 1 to 3 provide a common framework for studying different notions of discrimination in the literature and—as

[30] The legal notion of disparate impact is captured in the landmark Supreme Court decision *Griggs v. Duke Power Co.* (1971), which found that Duke Power's policy of requiring a high school diploma for within-company transfers was discriminatory because it disadvantaged Black employees who were qualified but lacked a degree, in part due to ongoing discrimination in secondary education. The Court noted that the degree requirement bore no relevance to an individual's ability to perform different jobs at the company. Notably, discrimination was found despite the policy being facially "race-blind": White and Black employees with the same educational background had the same ability to transfer jobs.

[31] Strack and Yang (forthcoming) characterize the set of signals that satisfy a 'privacy-preserving' property. Applying this characterization to our framework yields the set of signals that do not generate disparate impact for any action rule.

[32] An open debate is whether group-based differences in preferences that generate productivity or signal differences should be coded as discrimination. For example, gender socialization may lead female workers to refrain from asking for a raise, which then impacts their benchmark for future wage setting (e.g., Cook et al. 2021). The subsystem and qualification can be chosen to include or exclude such preference differences.

in the motivating example—can help interpret and integrate seemingly disparate findings.

Sources of Systemic Discrimination. Systemic discrimination can stem from an *informational* source, via group-based differences in the focal-node signal for workers who are equally productive at the focal node, or a *technological* source, via group-based differences in focal-node productivity for workers who are equally qualified at entry to the subsystem. The former captures the impact of other decisions on the ability to signal productivity, while the latter captures the impact of other decisions on the ability to build productivity.

Informational systemic discrimination can stem from *signal inflation*, where the signal is, on average, higher for one group than the other, conditional on focal node productivity. An example is when bail decisions depend on criminal records, and Black defendants have longer criminal records than White defendants with the same pretrial misconduct potential due to direct discrimination in other law enforcement interactions.[33] Other examples 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 (Agan et al. 2024). It can also arise from *differential screening*, where the manager has a more precise signal for one group than the other (Cornell and Welch 1996).[34] For example, a hiring process optimized for men may generate less informative signals for women (Pinkston 2003; Mocanu 2023).[35] Another example is how discrimination in access to past borrowing opportunities leads to differentially informative credit histories amongst borrowers with the same ability to repay (Bartik and Nelson 2024). See Appendix B for an empirical illustration of differential screening.

Similarly, technological systemic discrimination emerges when differences in productivity arise from discrimination at other nodes. For example, White workers might have more opportunities to build human capital than Black workers due to discrimination in access to training and skill development.[36] Alternatively, Black workers may respond to anticipated future direct discrimination by investing less in human capital (Lundberg and Startz 1983; Coate and Loury 1993).[37] Technological systemic discrimination also includes the type of

[33]For example, Pierson et al. (2020) show that Black individuals are more likely to be stopped by police and charged with a crime. Such criminal record disparities have also been shown to contribute to systemic discrimination in the labor market (Pager et al. 2009; Agan and Starr 2017).

[34]In Cornell and Welch (1996), minority job applicants receive fewer signal draws than majority applicants; there is no direct discrimination since applicants with the same number of draws are treated equally.

[35]Similarly, De Plevitz (2007) documents racial disparities in job screening stemming from using height-to-weight ratios calibrated with Anglo-Celtic data.

[36]For example, Gallen and Wasserman (2021) document 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.

[37]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^*).

"task-based" discrimination studied in Hurst et al. (2024), where racial barriers to task specialization generate group-based differences in choice of specialization. Importantly, our framework allows researchers to develop designs for bringing these concepts to data.[38]

Understanding the source of systemic discrimination is crucial for identifying an appropriate way to counteract it. For example, since signal inflation is a statistical bias in the signal, it can be offset via "reverse" direct discrimination in the action rule. In the case discussed earlier, if the bail judge is aware of how direct discrimination contributes to criminal record disparities, then she can account for it in her interpretation of criminal records. Hence, one potential policy response is ensuring awareness of signal inflation. In contrast, differential screening and technological systemic discrimination cannot be offset by simply adjusting the action rule without payoff consequences. To offset differential screening, more information needs to be collected for one group or the additional information needs to be ignored for the other group, creating a trade-off between equity and accuracy. Similarly, offsetting technological systemic discrimination requires additional investment in training for the disadvantaged group.

Relation to Statistical Discrimination. Our measure of systemic discrimination stems from group-based differences in the signal and productivity distributions, as with classical statistical (direct) discrimination. Statistical discrimination differs from systemic discrimination, however: it arises from the impact of these distributions on the action rule for a given signal realization, rather than 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 and statistical discrimination, and similarly for differences in the productivity distribution when $Y_i^0 \neq Y_i^*$ or $Y_i^0 \neq S_i^*$. In both cases, focusing only on direct discrimination can miss a key aspect of how group differences in these distributions contribute to action disparities.[39]

4 Identification

We now show how the framework can be brought to data with the iterated audit (IA) approach. We formalize the approach with a two-node subsystem with the reference qualifi-

[38]Consider, for example, a two-node system with a hiring focal node and a node corresponding to the acquirement of job-relevant human capital. A researcher studying this system with a constructive IA design (discussed below) could learn whether racial disparities in hiring are due to a *choice* made by workers to invest in human capital in anticipation of future discrimination (as in Coate and Loury 1993 and Figure 2(a)), or whether hiring disparities are due to direct discrimination in the opportunity to obtain human capital (as in Hurst et al. 2024 and Figure 1(a)).

[39]Unlike with statistical discrimination (e.g., Bordalo et al. 2019; Bohren et al. 2023) there is no scope for "inaccurate" systemic discrimination: only the true productivity and signal distributions contribute to systemic discrimination. However, inaccurate beliefs about these distributions can lead to inaccurate perceptions about the extent of systemic discrimination. This can affect action choices, and hence, total discrimination. For example, a mortgage assessor may incorrectly believe that a credit score is an unbiased signal of creditworthiness for groups, and therefore fail to adjust for its bias.

cation chosen to be the initial-node signal, $Y_i^0 = S_i^1$. We define $S^*(a^1, y^0)$ as the focal-node signal that is realized given an initial-node action of a^1 among those with $Y_i^0 = y^0$, such that $S_i^* = S^*(A_i^1, S_i^1)$. We assume that the researcher observes Y_i^0, A_i^1 and S_i^*.[40] We discuss at the end of this section how the approach extends to multi-node subsystems and other choices of reference qualification.

4.1 Treatment and Interaction Components

Measuring total and systemic discrimination 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 total and systemic discrimination is therefore identification of the interaction component.

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

$$\begin{aligned} \mu^*(y^0, g) &= E[A^*(g, S_i^*) \mid Y_i^0 = y^0, G_i = g] \\ &= A^*(g, S^*(A^1(g, y^0), y^0)) \qquad (7) \\ \tilde{\mu}^*(y^0, g) &= E[A^*(g', S_i^*) \mid Y_i^0 = y^0 G_i = g] \\ &= A^*(g', S^*(A^1(g, y^0), y^0)), \qquad (8) \end{aligned}$$

recalling that $A^*(g, s)$ and $A^1(g, s)$ are the action rules at the focal and initial node, respectively. 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 (i.e., how $A^*(g, S^*(a^1, y^0))$ depends on a^1). 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

[40] For ease of notation, we implicitly assume that S_i^1 is the only difference across workers at entry to the subsystem and that the focal-node signal only depends on S_i^1 and A_i^1. In practice, workers could also differ along other components of the focal-node signal S_i^* that are independent of A_i^1, in which case the reference qualification would also include these components. If the researcher did not observe these other components, then ensuring that their distributions are similar across groups (as in an audit/correspondence study) identifies our three measures of discrimination.

position (the focal node). Equations (7) and (8) show that total and 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 iterated audit approaches to estimate the interaction component: a *constructive* approach, which separately estimates the impact of each possible action at the other node on the focal node action, and an *experimental* approach, which directly measures the interaction component by simulating the impact of other-node action distributions.

4.2 The Constructive IA Approach

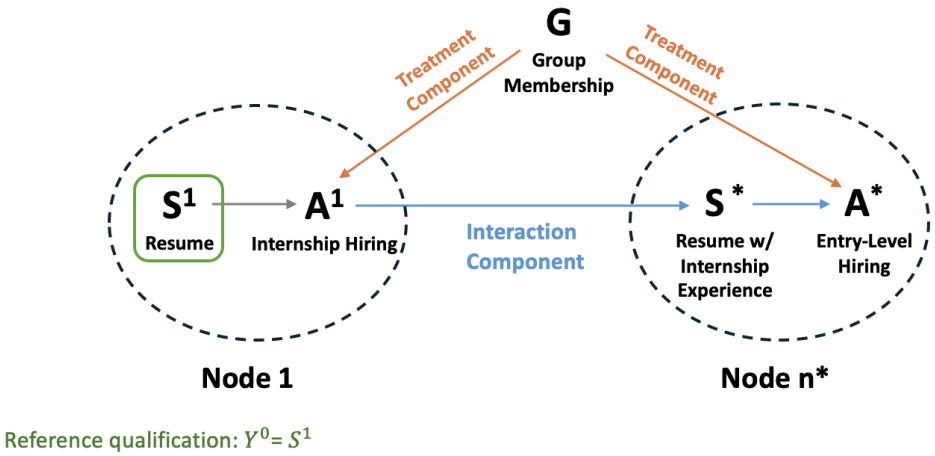
The first IA approach separately estimates the effect of each possible action at non-focal nodes on the focal-node signal, along with the effect of each possible focal-node signal realization on the focal-node action. In our two-node example, this means learning how $S^*(a^1, y^0)$ depends on a^1 and how $A^*(g, s)$ depends on s. With these, a researcher can *construct* an estimate of the interaction component: e.g. how $A^*(g, S^*(a^1, y^0))$ depends on a^1.

To make this constructive IA approach concrete, consider again the internship and entry-level hiring example. A researcher could first learn how past internship experience a^1 translates to the signal $S^*(a^1, y^0)$ that a hiring manager observes when evaluating candidates for the entry-level job; e.g., she could see that candidates with internship experience list it on resumes sent to hiring managers. The researcher could then learn how hiring manager actions depend on this experience; 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.

Figure 3 illustrates the constructive IA design for this example. The treatment component is given by estimates of how group membership directly affects internship and entry-level hiring (the orange arrows). The interaction component is given by estimates of how internship hiring impacts entry-level hiring through resume internship experience (the blue arrows). The combination of these estimated effects identifies total discrimination in entry-level hiring. Removing the impact of the orange arrow on the right, which captures direct discrimination in entry-level hiring, identifies systemic discrimination.

This example also highlights the importance of measuring the interaction component and why conventional audit study estimates of the treatment component are not enough to identify systemic or total discrimination. If the level of internship experience is critical

FIGURE 3. Constructive IA Design in Entry Level Hiring



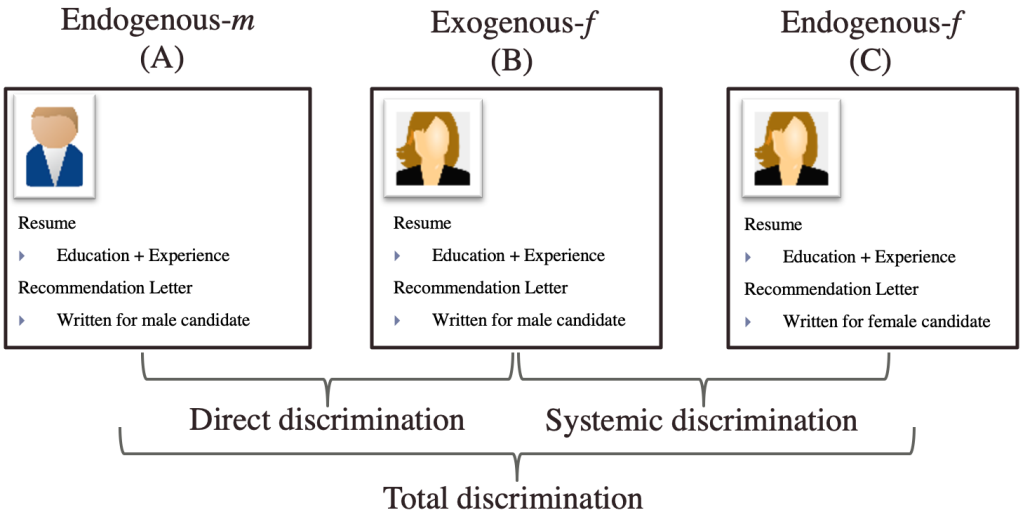
for future success in hiring, then minimal direct discrimination in internship hiring could lead to large systemic and total discrimination in entry-level hiring through the interaction component. More generally, nonlinearities in how actions at non-focal nodes affect focal-node actions means that conventional audit study analyses of direct discrimination at multiple nodes will fail to reveal the total extent of discrimination arising from interactions across nodes. Conversely, decision-makers might act to undo perceived direct discrimination in the signals they observe, such that direct discrimination at non-focal nodes does not translate to total discrimination at the focal node. For example, entry-level hiring managers may 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, it may translate to minimal total discrimination in entry-level hiring.

While conceptually straightforward, the constructive IA approach may be challenging to implement when action or signal spaces are high-dimensional or otherwise complex. Suppose, for example, that workers apply to the entry-level position with both a resume and recommendation letter written by the internship supervisor. This supervisor discriminates in the language and tone of the letters for equally qualified workers. Here the set of signals (i.e., possible recommendation letters) is high-dimensional, making it hard to estimate how $A^*(g, s)$ depends on s. To address such issues, we turn to the second IA approach.

4.3 The Experimental IA Approach

The second IA approach directly measures systemic and total discrimination by simulating the distribution of focal-node signals as impacted by group membership through the actions at other nodes, and measuring focal-node actions given the simulated signal distributions. In our two-node example, measuring $\Delta^S(y^0, w)$ and $\Delta^T(y^0)$ means generating draws of $\tilde{S}_i^* \mid G_i = g, Y_i^0 = y^0$ for $g \in \{w, b\}$ and measuring $A^*(w, \tilde{S}_i^*(w, y^0))$, $A^*(b, \tilde{S}_i^*(b, y^0))$ and

FIGURE 4. Experimental IA Design in Entry-Level Hiring



$A^*(w, \tilde{S}_i^*(b, y^0))$ to *experimentally* measure $\mu^*(y^0, w)$, $\mu^*(y^0, b)$ and $\tilde{\mu}^*(y^0, b)$.[41]

To make this approach concrete, consider a modified version of the previous example where entry-level hiring managers observe recommendation letters from the internship supervisors. A researcher could obtain a set of recommendation letters from individuals of each group. She could then generate entry-level job applications by drawing letters from these sets and attaching a salient signal of group membership—both the same group as the individual that the letter was written for, g, and the other group, g'. The average entry-level hiring manager action for these simulated applications identifies $\mu^*(y^0, g)$ when the application and letter were both assigned group g, and $\tilde{\mu}^*(y^0, g)$ when the application was assigned to group g' but the letter was written for an individual from group g. Together, these estimates identify systemic and total discrimination.

Figure 4 illustrates the experimental IA design for this example, focusing on gender discrimination. Hiring managers are presented with applications from one of three groups which allow the researcher to decompose total discrimination into average direct discrimination with respect to the male signal distribution and systemic discrimination under the female action rule. The first two groups, termed the *endogenous-m* and *exogenous-f* groups, are similar to those used in a standard correspondence or audit study: the resumes and recommendation letters are identical except for a signal of gender (e.g., a distinctively male or female name). Here, the recommendation letters are drawn from the set of male letters. The third group of applicants, termed the *endogenous-f* group, has resumes with distinctively female names but recommendation letters drawn from the set of female letters. Comparing the hiring actions of endogenous-m vs. exogenous-f applications identifies average direct dis-

[41] Analogously, using these simulated distributions to measure $A^*(w, \tilde{S}_i^*(w, y^0))$, $A^*(w, \tilde{S}_i^*(w, y^0))$ and $A^*(w, \tilde{S}_i^*(b, y^0))$ identifies $\Delta^S(y^0, b)$ and $\Delta^T(y^0)$.

crimination since it compares identical resumes and letters, as in a standard correspondence study, while comparing the actions of exogenous-f vs. endogenous-f applications identifies systemic discrimination. Finally, comparing endogenous-m vs. endogenous-f applications identifies 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 behind systemic discrimination may be harder to infer—which can be important for forming appropriate policy responses. Still, by capturing the interaction component, the experimental IA approach will also account for any nonlinearities or offsetting behavior that might obscure the impact of systemic discrimination if one were to try to infer it from accumulated estimates of direct discrimination at multiple nodes.

4.4 Extensions

It is straightforward to extend both IA approaches to subsystems with multiple nodes and to other observed reference qualifications. The treatment component would again be estimated by the direct effects of group membership on actions at the focal node and all other nodes. In the constructive IA approach, the researcher would learn how the focal-node signal $S^*(a^1, ..., a^N, y^0)$ depends on all other-node actions $(a^1, \ldots, a^N)$, as well as how the focal-node action rule $A^*(g, s)$ depends on s. These estimates can be used to construct the interaction component, i.e., the dependence of $A^*(g, S^*(a^1, ..., a^N, y^0))$ on $(a^1, ..., a^N)$, via generalized versions of Equations (7) and (8). In the experimental IA approach, the researcher would simulate the conditional-on-Y_i^0 distribution of focal-node signals, as impacted by group membership through the actions at all other nodes, and again measure focal-node actions given the simulated signal distributions.

In subsystems with many nodes, an experimental IA may be more tractable than an constructive IA as it avoids the need to estimate the effect of each possible action $a^n \in \mathcal{A}^n$ for $n = 1, ..., N$ on the focal-node signal. On the other hand, the constructive IA approach may be more useful for studying mechanisms, identifying the nodes that generate the biggest systemic disparities, or forming appropriate policy responses, as it separates out each indirect effect that gives rise to systemic discrimination. These tradeoffs mirror those from the discussion of complex or high-dimensional signals in the two-node case.

Extensions to partially observed or unobserved Y_i^0 are likely more involved. Suppose, for example, the reference qualification is chosen to be focal-node productivity: $Y_i^0 = Y_i^*$. In some settings, productivity may be observed only selectively: e.g., hired workers reveal their productivity by the number of units produced on the job, which is unobserved among unhired workers. The interaction component—and hence systemic and total discrimination—will typically be only partially identified in this case, though additional quasi-experimental

variation in the outcome selection mechanism can bridge this gap (see, e.g., Arnold et al. 2022; Baron et al. 2024; Rambachan et al. 2024). In other settings, Y_i^0 may not even be selectively observed and must be proxied by other observables. Here, frameworks like those in Altonji et al. (2005) or Oster (2017) may be used to gauge the sufficiency of the proxies for capturing systemic and total discrimination. As before, we expect the experimental IA approach to be more tractable in these extensions by reducing the number of conditional-on-Y_i^0 objects to be estimated.

5 Constructive IA Application

Our first field experiment uses the constructive IA approach to explore the impact of racial discrimination in initial work experience on later hiring outcomes.[42] We focus on a two-node subsystem in which the first node generates an action (experience) that is part of the signal at the second, focal node (a later hiring decision). At the first node, Black and White workers apply to a set of jobs with no prior work experience. They are either hired for one of the jobs or not. They then apply for a job at the focal node either with or without prior experience from the first node, depending on whether or not they were hired. Direct discrimination in hiring inexperienced workers can thus generate systemic discrimination in subsequent hiring due to racial disparities in work experience. We use a correspondence study to estimate direct discrimination for experienced and inexperienced workers, and then use additional data to construct the interaction component and estimate systemic and total discrimination.

5.1 Methods

We first measured direct discrimination in callback rates for inexperienced versus experienced applicants at a national sample of automotive firms. We targeted entry-level positions in the automotive sector because it had previously been shown to exhibit sizable racial discrimination in callback rates (Kline et al. 2022).[43] Following a standard correspondence study design, we generated sets of fictitious resumes that varied in the applicant's race and non-race attributes (e.g., their address, high school). Race was randomly varied through the applicant's name, which was either distinctively White or Black following the literature (e.g., Bertrand and Mullainathan 2004). Job experience was also randomly varied by whether the resume included relevant experience, corresponding to one prior job in the automotive sector, or not. The resumes were submitted to online job postings and subsequent callbacks were recorded. See Appendix C.1.1-Appendix C.1.4 for details on the experimental design.

[42] See https://aspredicted.org/5XT_YPY for pre-registration materials. Per the pre-registration, we first ran a pilot to determine the feasibility of the experimental design. There were no qualitative differences in any of the comparisons between the pilot and full study. As outlined in the document, we pool the data for the main analysis.

[43] We used the same sample of automotive firms as in Kline et al. (2022). Notably, our experiment was completed prior to the public identification of those firms in Kline et al. (2024).

To estimate the likelihood that a callback translated to a job offer, and hence experience, we obtained a measure of the typical conversion rate for callbacks. Specifically, we used a recruiting agency to survey 107 hiring managers with experience in evaluating applicants for entry-level jobs in the automotive industry. We elicited the managers' estimated probability that a callback would lead to a job offer as a function of the applicant's previous experience and race. See Appendix C.1.5 for details on the survey.

Mapping this data to our framework, consider the two-node subsystem $\mathcal{N}^0 = \{1, n^*\}$ where node 1 corresponds to an initial job search and focal node n^* corresponds to a hiring decision in a subsequent job search. At node 1, worker i applies to J_i jobs with resume S_i^1 (which does not contain any relevant experience) and receives hiring decision $A_i^1 = (A_i^{1,1}, ..., A_i^{1,J_i})$, where $A_i^{1,j} = A^{1,j}(G_i, S_i^1)$ denotes whether or not the worker is offered job j. We assume that a worker can accept at most one job offer and if the worker receives at least one offer, then she accepts it and it translates to a line of experience. Therefore, this first-node hiring action A_i^1 affects the worker's second-node resume S_i^* (the node-n^* signal) by either adding a line of relevant experience or not to the first-node resume S_i^1, depending on whether the worker obtained at least one job offer or not. At node n^*, worker i applies to a job with resume S_i^* and receives hiring decision $A_i^* = A^*(G_i, S_i^*)$, which corresponds to whether or not she is offered the job. The dependence of A_i^* on A_i^1 via S_i^* captures the link between first-node hiring and second-node hiring: direct discrimination at the first node ($A^1(b, S_i^1) \neq A^1(w, S_i^1)$) can translate to different rates of experience, and hence, systemic discrimination in subsequent hiring. We set the reference qualification as the first-node resume, $Y_i^0 = S_i^1$, to isolate the systemic impact of direct discrimination in hiring inexperienced workers.

The data described above generates an estimate of the treatment component in this subsystem, i.e., direct discrimination at each node and signal. The data on hiring conversion rates allows us to translate callback rates to hiring rates at each node—for node 1, the average hiring rate $A_i^{1,j}$ across workers with no experience, and for node n^*, the average hiring rate A_i^* across workers with each level of experience. Specifically, multiplying the average callback rate and average callback conversion rate for a given race and experience level yields an estimate of the corresponding job offer rate. Direct discrimination in first-node hiring then corresponds to the difference between the job offer rate for White versus Black workers with no relevant experience. This is also the estimate of direct discrimination in second-node hiring for workers who failed to gain experience at the first node. Direct discrimination in second-node hiring for experienced workers is the analogous difference for White versus Black workers with relevant experience from the first node.

To estimate the interaction component, i.e., how direct discrimination in first-node hiring indirectly affects second-node hiring through listed job experience, we also need a measure of the likelihood of gaining experience at the first node. This depends on the number of

jobs the applicant applies to. We proxy this number with data on the number of local job openings (i.e., local labor market thickness). Specifically, we scraped the average number of entry-level job openings at automotive firms over a week for each municipality included in the experiment. See Appendix C.1.6 for details. This data, together with the job offer rate estimates described above, yields an estimate of the probability of becoming experienced at the second node by race.[44]

We combine these components to construct estimates of systemic and total discrimination, following Section 4.2. We average across qualification levels $Y_i^0 = S_i^1$ (i.e., first-node resumes), as the design imposes the same qualification distribution across race.

5.2 Results

We first analyze the callback rates in the experiment, as reported in Table 1. Column 1 shows sizable direct discrimination in callback rates against Black applicants. Overall, 28.3% of White applicants received a callback compared to 20.0% of Black applicants—a significant 8.3 percentage point gap ($p < 0.01$). Columns 2 and 3 further show substantial heterogeneity with respect to prior job experience. Among resumes without experience, callback rates were 25.6% for White applicants and 13.2% for Black applicants, resulting in a significant 12.4 percentage point disparity ($p < 0.01$). This gap is larger than callback disparities found in previous studies, which, to the best of our knowledge, included previous work experience on the resumes (Bertrand and Mullainathan 2004; Nunley et al. 2015; Deming et al. 2016). Indeed, experience shrank the racial disparity to 4.2 percentage points, with 31.1% of White applicants and 26.9% of Black applicants receiving a callback. While not statistically significant ($p = 0.20$), this disparity roughly matches the 5.3 percentage point gap ($p < 0.01$) that Kline et al. (2024)—who only used applications with experience—find for the same types of automotive sector jobs. Column 4 summarizes the return to experience: among all applicants, experienced resumes received callbacks at a 9.5 percentage point (49%) higher rate than inexperienced resumes ($p < 0.01$).

Turning to callback conversion rates, hiring managers reported that, on average, 55% of applicants without experience received a job offer, compared to 71% of applicants with experience. This difference is statistically significant ($p < 0.01$). The racial disparity in reported callback conversion rates—approximately 3 percentage points—was not significant. These relatively high conversion rates suggest that the racial disparities in callback rates will translate into meaningful disparities in job offers.

Finally, local market thickness was moderate and varied across municipalities: the 25^{th},

[44]Specifically, we assume the probability of being offered each job is independently distributed for inexperienced workers conditional on race. Letting p_g denote this probability for race g, the probability of receiving no job offers is $(1 - p_g)^J$, where J corresponds to the number of local job openings, and the probability of gaining experience is $1 - (1 - p_g)^J$.

TABLE 1. Constructive IA Application: Effects of Race and Experience on Callback Rates

	(1) Full Sample	(2) Inexperienced	(3) Experienced	(4) Full Sample
Black	-0.083***	-0.124***	-0.042	
	(0.022)	(0.030)	(0.033)	
Experienced				0.095***
				(0.021)
Constant	0.283***	0.256***	0.311***	0.194***
	(0.023)	(0.028)	(0.030)	(0.020)
Observations	1,001	500	501	1,001

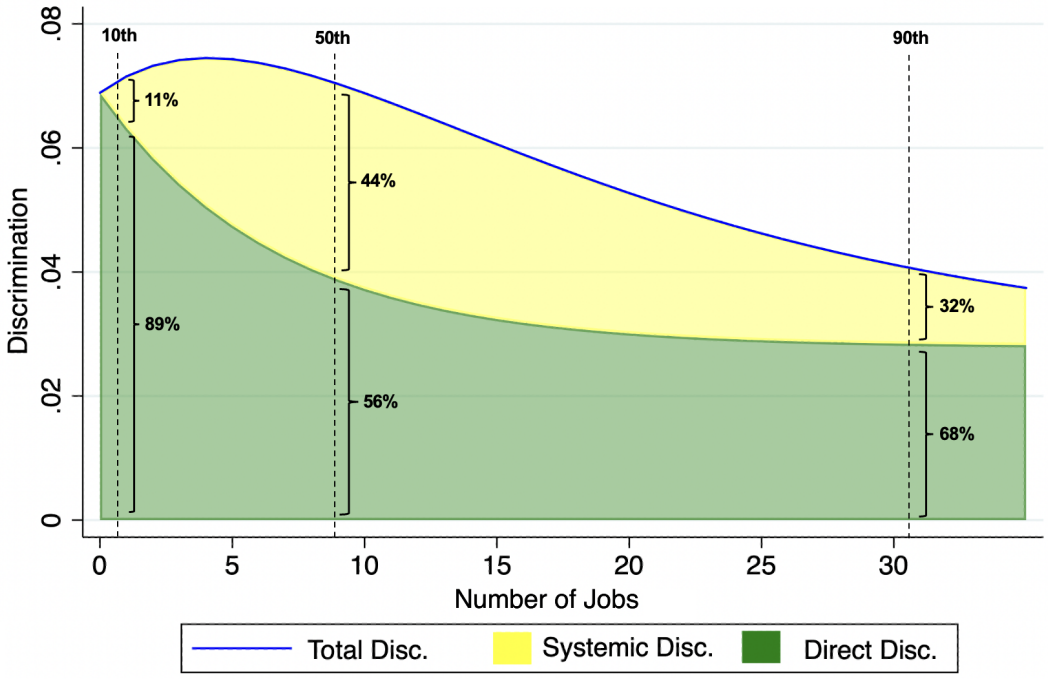
Notes: This table reports estimated coefficients from regressing a callback indicator on either an indicator for a resume having a distinctively Black name or a resume having prior relevant job experience. Standard errors, clustered at the posting level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

50^{th}, and 75^{th} percentiles of number of job postings were 5, 8, and 18, respectively. The moderate number of postings suggests that the relatively large racial disparity in job offer rates for inexperienced workers will lead to notable disparities in second-node experience, since applicants do not have a large number of chances to receive an offer. Moreover, the variability in local market thickness suggests that the systemic discrimination stemming from this experience disparity will vary meaningfully across regions.

Figure 5 shows estimates of total, systemic, and direct discrimination at the second (focal) node as a function of the number of first-node job openings. In markets with a median thickness of 8 openings, White applicants are around 7.5 percentage points more likely to be hired than Black applicants, after having had a chance to secure prior job experience. More than half of this total discrimination is due to systemic discrimination against Black applicants, who are less likely to have valuable prior experience due to direct discrimination in first-node hiring. The remainder is due to direct discrimination in second-node hiring, where White applicants are more likely to be hired than Black applicants with the same experience. A standard correspondence study—which conditions on experience—would therefore miss nearly half of the total discrimination in this subsystem.

These estimates demonstrate that market conditions (here, local labor market thickness) can play a key role in amplifying or mitigating the impact of past discrimination. When the market is very thin or thick, direct discrimination in first-node hiring will not generate large differences in experience. Intuitively, if the market is thin with very few openings, then few applicants gain experience, regardless of race. Most enter the second node without experience, and direct discrimination towards inexperienced Black workers drives total discrimination. At the other extreme, in thick markets with many openings, applicants

FIGURE 5. Constructive IA Application: Discrimination Decomposition



Note: This figure shows total discrimination $\Delta^T(y^0)$, average direct discrimination $\Delta^D(y^0, w)$, and systemic discrimination $\Delta^S(y^0, b)$ as a function of local market thickness, averaged across qualification levels y^0.

have many chances to be hired and gain experience. Therefore, the hiring decision for any one opening has little impact on future experience—most applicants land a job and enter the second node with experience. Hence direct discrimination towards experienced Black workers drives total discrimination. In contrast, in markets of intermediate thickness, direct discrimination in first-node hiring can translate to large differences in experience. In our data, systemic discrimination is 11% of total discrimination in thin markets with only one job posting (10^{th} percentile) and 32% in thick markets with 31 jobs (90^{th} percentile), but increases to 44% in intermediate markets with 8 jobs (50^{th} percentile).

From both a conceptual and policy perspective, this analysis suggests that market thickness is an important factor when interpreting disparities in callback or hiring rates. First, it provides a broader lens for interpreting prior results on callback and interview rate disparities by showing how they translate into labor market outcomes. Second, the analysis informs the efficacy of potential interventions aimed at reducing systemic discrimination: targeting initial disparities is most important in markets of intermediate thickness.

It is worth noting that, while we illustrate the approach in a simple two-node system, our findings suggest that systemic discrimination may be even more pronounced when work experience is affected by direct discrimination across more than two hiring nodes. In this case, conventional correspondence studies could miss an even larger share of total discrimination

by conditioning on all prior work experience.

6 Experimental IA Application

Recent studies find 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 (Trix and Psenka 2003; Schmader et al. 2007). But whether and how such differences in language contribute to disparities in labor market outcomes is not obvious. For example, hiring managers may focus on 'hard' information and ignore language differences; alternatively, the differences may lie in language that is not relevant for the hiring decision.

We use the experimental IA approach to explore the potential impact of language on hiring outcomes in a labor market for STEM positions. Building on the example discussed in Section 4.3, we focus on a two-node hiring subsystem in which the first node generates an action (the recommendation letter) that is part of the signal at the second focal node (the hiring decision).[45] At the first node, men and women with similar resumes are evaluated and receive a recommendation letter. They then apply for a job at the second (focal) node with these letters and their resumes. Direct discrimination in the content and style of the recommendation letter (i.e., gendered language and tone) can thus generate systemic discrimination in hiring. Applying the experimental IA approach, we simulate letters for male versus female applicants and elicit hiring outcomes and wage offers for the three groups outlined in Figure 4. This yields estimates of systemic and total discrimination in hiring due to gender disparities in recommendation letters. This experimental design avoids the challenges of estimating the interaction component in the constructive IA approach, which would be prohibitive given the high dimensionality of the recommendation letter signal.

6.1 Methods

For the first node, we used a large language model (LLM) to generate a set of recommendation letters for fictitious job applicants with recent college degrees in science, technology, engineering, or mathematics (STEM). LLMs have been shown to generate gendered language that matches differences in letters written by humans (Wan et al. 2023). We generated a set of resumes that varied in the applicant's gender, as signaled through name (distinctively male or female). These resumes were otherwise identical in terms of relevant education and experience. We then prompted the LLM to generate a recommendation letter for each resume. This approach lets us hold fixed the relevant qualifications of candidates while replicating the gendered language and tone employed by recommendation letter writers. Indeed, the

[45] See https://aspredicted.org/w4g6-kzgq.pdf for pre-registration materials. We pre-registered three sets of materials based on the design in Figure 4. As outlined below, we ran a fourth group to check robustness to the alternative decomposition.

letters in our study exhibited similar gender differences—on dimensions such as individual agency, leadership ability, and communality—as previously documented for both LLMs and humans (Trix and Psenka 2003; Schmader et al. 2007).[46] See Appendix C.2 for details.

For the second (focal) node, we used a professional survey targeting agency to recruit 396 hiring managers. These managers had experience in evaluating applicants for STEM jobs that typically require a recommendation letter. Each manager observed the resumes and recommendation letters for a random set of two applicants. They then reported the likelihood that they would recommend each applicant to the next stage of the recruitment process (on a scale of 1 to 10) and each applicant's expected hourly wage should they be hired. These reports were incentivized using a similar methodology to Kessler et al. (2019): the managers were aware that the applicants were fictitious, but the attributes (e.g., prior work experience) could be matched to actual applicants with similar attributes. Managers were shown matched real applicants based on the fictitious applicants that they rated highly.[47]

Applicants were drawn from four groups, mirroring the groups illustrated in Figure 4. The endogenous-m and exogenous-f groups were similar to those which would be used in a standard correspondence or audit study: the resumes and recommendation letters were identical except for using distinctively male or female names and pronouns. Specifically, the resumes differed only in the candidate's name, and the recommendation letters were generated from this resume with a male name; for the exogenous-f group, the name and pronouns were then changed to be distinctively female. The endogenous-f group had resumes with distinctively female names, but differed from the exogenous-f and endogenous-m groups in that the recommendation letters were generated from the resume with a female name. The fourth group—exogenous-m—had resumes with distinctively male names and recommendation letters generated from the resume with a female name. Adding this group enabled us to also measure systemic discrimination under the male action rule.

Mapping this data to our framework, given a two-node subsystem $\mathcal{N}^0 = \{1, n^*\}$, node 1 corresponds to obtaining a recommendation letter and focal-node n^* corresponds to a hiring decision. The first-node action $A_i^1 = A^1(G_i, S_i^1)$ corresponds to the letter for worker i of gender G_i based on resume S_i^1. This action is part of the focal-node signal $S_i^* = (A_i^1, S_i^1)$, which consists of the recommendation letter and resume. Direct discrimination at the first node ($A^1(f, S_i^1) \neq A^1(m, S_i^1)$, where we set $G_i \in \{m, f\}$ to denote male and female) corresponds to using systematically different language when writing letters for male versus female workers with the same resume. This direct discrimination can generate systemic

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

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

discrimination in focal-node hiring and wage actions $A_i^* = A^*(G_i, S_i^*)$, given the dependence of S_i^* on A_i^1. We set the reference qualification as the resume, $Y_i^0 = S_i^1$, to isolate the systemic impact of direct discrimination in letter writing.

Critically, relative to the study in Section 5, the component of signal S_i^* generated in the subsystem (i.e., the letter) is high-dimensional. This precludes measuring the impact of each signal realization, which would be required for estimating the interaction component in the constructive IA approach. Instead, the experimental approach constructs measures of systemic and total discrimination from a comparison of the four groups outlined above. Specifically, comparing the hiring outcomes of the endogenous-m vs. exogenous-f groups identifies average direct discrimination under the male signal distribution, comparing exogenous-f vs. endogenous-f identifies systemic discrimination under the female action rule, and comparing endogenous-m vs. endogenous-f identifies total discrimination. Analogously, the exogenous-m group together with endogenous-m and endogenous-f allows us to measure systemic discrimination under the male action rule. The analysis that follows focuses on the first case, with the second case presented below for robustness.

6.2 Results

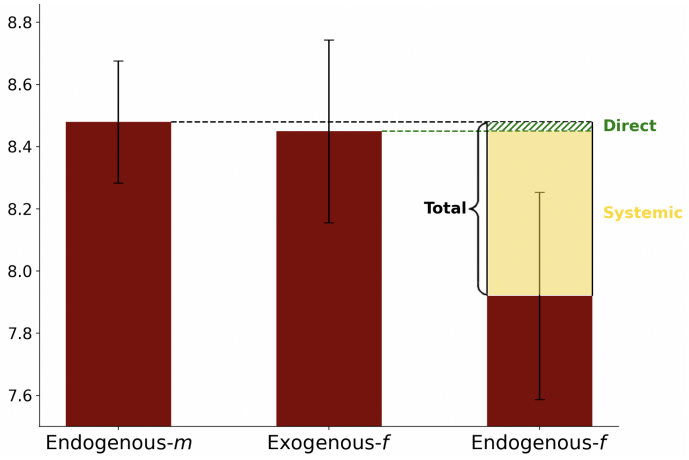
We found substantial total discrimination against female applicants. Figure 6 shows that the endogenous-f group had significantly lower average hiring likelihood and prospective wage than the endogenous-m group, corresponding to roughly 29% and 31% of the respective outcome standard deviations ($p < 0.01$ for both outcomes). In contrast, we found small and insignificant levels of direct discrimination. The average hiring likelihood for the exogenous-f group was only slightly below that of the endogenous-m group, while exogenous-f had a slightly *higher* average prospective wage (though this gap is not statistically significant). As discussed below, these results are consistent with recent work finding minimal direct gender discrimination in audit and correspondence studies.

As shown in Figure 6, this significant total discrimination is driven by systemic discrimination that stems from gendered differences in the language used in the recommendation letters. The endogenous-f group—who had letters written for female applicants—had a substantially lower hiring likelihood and prospective wage than the exogenous-f group—who had letters written for male applicants. These gaps represent 27% and 43% of the respective outcome's standard deviation and the vast majority (or more) of total discrimination.[48]

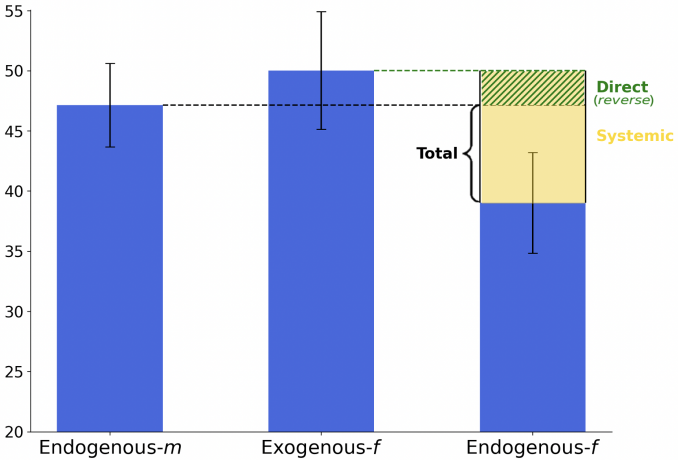
To demonstrate the utility of the experimental IA approach in this setting, we consider an alternative analysis that uses low-dimensional statistics to summarize the language of the high-dimensional recommendation letter signal. Specifically, we follow Kaplan et al. (2024) in

[48]The alternative decomposition yielded similar findings. Both the hiring likelihood and prospective wage for the exogenous-m group (7.36 and 38.85, respectively) was similar to the endogenous-f group, indicating little direct discrimination. Thus, this decomposition also shows that most total discrimination is systemic.

FIGURE 6. Experimental IA Application: Results



(a) Hiring Likelihood



(b) Prospective Wage

Note: This figure shows average hiring likelihoods and prospective wages for the three experimental IA groups, as well as the resulting estimates of total, direct, and systemic discrimination. The error bars report 95% confidence intervals.

summarizing letter language across the four dimensions that have been found to be strongly gendered in recommendation letters (Schmader et al. 2007). These indices measure language associated with ability, agency, leadership, and being a standout job candidate.

If these indices successfully captured the key aspects of language responsible for generating the observed hiring and wage disparities, then one could potentially use the constructive IA approach with text data via such summary statistics. Table 2 shows that this would be challenging here: controlling for this set of indices has little impact on the measured gender gap in hiring likelihoods or prospective wages between the endogenous-m and endogenous-f groups. In the case of the hiring likelihood outcome, the controls lead to a small decrease in the gender gap; in the case of prospective wages, the gender gap actually increases. In both

TABLE 2. Experimental IA Application: Effects of Gender and Language Constructs

	Hiring Likelihood		Prospective Wage	
	(1)	(2)	(3)	(4)
Female	-0.55***	-0.50**	-8.11***	-12.80***
	(0.16)	(0.23)	(1.73)	(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)
Observations	441	441	441	441

Notes: This table reports estimated coefficients from regressing either hiring likelihood or prospective wage on a female indicator and four indices summarizing recommendation letter language. Standard errors, clustered at the hiring manager level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

cases, none of the four indices are statistically significant. In other words, the low-dimensional summaries of language account for a minimal share of what drives systemic discrimination in this setting.[49] This highlights the difficulty of examining the impact of complex signals such as text data using the constructive IA approach. In our setting, without the experimental IA design, a researcher could erroneously conclude that gendered language disparities do not contribute to hiring discrimination.

Substantively, these findings speak to recent work which argues that direct discrimination is unlikely to fully explain gender disparities in the labor market. For example, a meta-analysis of correspondence studies in Schaerer et al. (2023) found little discrimination overall when non-group attributes (i.e., the signal) for male and female candidates are held fixed (although there was substantial heterogeneity across employers, with some favoring women and others favoring men; see also Kline et al. 2022). Our findings suggest that, by controlling for non-group attributes, such estimates may miss a substantial degree of gender discrimination from "soft" signals in verbal recommendations. These complex soft signals may ultimately be responsible for driving a substantial share of observed gender disparities in labor markets. This provides further evidence for the importance of relying on "hard" signals in the evaluation process.[50]

[49]For example, systemic discrimination could arise from complex interactions between different components of the letter or difficult-to-quantify aspects of the text.

[50]See Amer et al. (2024) for evidence that personal interaction—which has a similarly complex signal space as recommendation letters—may be responsible for gender bias in the technology industry.

7 Conclusion

Several large literatures, both within and outside of economics, emphasize the importance of systemic factors in driving group-based disparities. But empirical analyses largely focus on direct discrimination as a function of group identity itself. We bridge this gap by developing new theoretical and empirical tools to study and measure systemic discrimination. Our general framework formalizes a notion of systemic discrimination as arising from discrimination at other points in a system, and defines a measure of total discrimination that accounts for both direct and systemic components. We map our framework to past work on discrimination, showing that our definition of systemic discrimination unearths forces in these existing models that may have been overlooked. We then develop new empirical tools—the constructive and experimental IA designs—to identify these components in the data. Our applications demonstrate the presence of systemic discrimination in two important labor market settings, and show how conventional methods of studying direct discrimination can lead to underestimates of the total impact of discrimination in a given setting.

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 the 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. 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 was found in the landmark *Griggs v. Duke Power Co. (1971)* case. The development of robust econometric methods for measuring systemic and total discrimination can be a powerful complement to existing regulatory tools in such settings.[51] Economic models of systemic discrimination can aid the interpretation of these methods, by enriching policymakers' understanding of interactions over time and across different domains.

[51] 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 discrimination 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).

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.

Abramitzky, R., J. Conway, R. Mill, and L. C. Stein (2023): "The Gendered Impacts of Perceived Skin Tone: Evidence from African American Siblings in 1870–1940," Working Paper 31016, National Bureau of Economic Research.

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 (2024): "Salary History and Employer Demand: Evidence from a Two-Sided Audit," *American Economic Journal: Applied Economics*, forthcoming.

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 (2024): "Jim Crow and Black Economic Progress after Slavery," *The Quarterly Journal of Economics*, 139, 2279–2330.

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

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

Amer, A., A. C. Craig, and C. Van Effenterre (2024): "Decoding Gender Bias: The Role of Personal Interaction," IZA Discussion Paper 17077, Institute of Labor Economics.

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*, 112, 2992–3038.

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.

Avery, M., A. Leibbrandt, and J. Vecci (2023): "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in

Tech," Monash Economics Working Papers 2023-09, Monash University, Department of Economics.

Baron, E. J., J. J. Doyle Jr, N. Emanuel, P. Hull, and J. P. Ryan (2024): "Discrimination in Multiphase Systems: Evidence from Child Protection," *The Quarterly Journal of Economics*, 139, 1611–1664.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2024): "Explicit and Implicit Belief-Based Gender Discrimination: a Hiring Experiment," *Management Science*, forthcoming.

Bartik, A. W. and S. T. Nelson (2024): "Deleting a Signal: Evidence from Pre-Employment Credit Checks," *The Review of Economics and Statistics*, 1–47.

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

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2021): "Fairness in Criminal Justice Risk Assessments: The State of the Art," *Sociological Methods & Research*, 50, 1–42.

Bertrand, M. and E. Duflo (2017): "Field Experiments on Discrimination," in *Handbook of Economic Field Experiments*, ed. by A. V. Banerjee and E. Duflo, North-Holland, vol. 1, 309–393.

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

Blank, R. M. (2005): "Tracing the economic impact of cumulative discrimination," *American Economic Review*, 95, 99–103.

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 (2023): "Inaccurate Statistical Discrimination: An Identification Problem," *Review of Economics and Statistics*, 1–45.

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.

Bursztyn, L., T. Fujiwara, and A. Pallais (2017): "'Acting Wife': Marriage Market Incentives and Labor Market Incentives," *American Economic Review*, 107, 3288–3319.

Cain, G. G. (1986): "The Economic Analysis of Labor Market Discrimination: A Survey," in *Handbook of Labor Economics*, Elsevier, vol. 1, 693–785.

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.

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.

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.

Davison, H. K. and M. J. Burke (2000): "Sex Discrimination in Simulated Employment Contexts: A Meta-analytic Investigation," *Journal of Vocational Behavior*, 56, 225–248.

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," Working Paper 30367, National Bureau of Economic Research.

Deming, D. J., N. Yuchtman, A. Abulafi, C. Goldin, and L. F. Katz (2016): "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study," *American Economic Review*, 106, 778–806.

Derenoncourt, E., C. H. Kim, M. Kuhn, and M. Schularick (2024): "Wealth of

two nations: The US racial wealth gap, 1860–2020," *The Quarterly Journal of Economics*, 139, 693–750.

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. (1977): "Indirect Institutionalized Discrimination: A Typological and Policy Analysis," *American Politics Quarterly*, 5, 177–200.

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," in *The Handbook of Social Psychology*, ed. by D. T. Gilbert, S. T. Fiske, and G. Lindzey, McGraw-Hill, 357–411, 4th ed.

Fryer, R. G. and S. D. Levitt (2004): "The Causes and Consequences of Distinctively Black Names," *The Quarterly Journal of Economics*, 119, 767–805.

Gaddis, S. M. (2017): "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies," *Sociological Science*, 4, 469–489.

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," in *Proceedings of the Agricultural and Applied Economics Association (AAEA) 2022 Annual Meeting*.

Gebru, T. (2020): "Race and Gender," in *The Oxford Handbook of Ethics of AI*, Oxford University Press, 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 wp514, University of Chile, Department of Economics.

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," in *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?" Working Paper.

Hill, R. B. (1988): "Structural Discrimination: The Unintended Consequences of Institutional Processes," in *Surveying Social Life: Papers in Honor of Herbert H. Hyman*, ed. by H. J. O'Gorman, Wesleyan University Press, 353–375.

Hübert, R. and A. T. Little (2023): "A Behavioral Theory of Discrimination in Policing," *The Economic Journal*, 133, 2828–2843.

Hull, P. (2021): "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making," Working Paper 28503, National Bureau of Economic Research.

Hurst, E., Y. Rubinstein, and K. Shimizu (2024): "Task-Based Discrimination," *American Economic Review*, 114, 1723–1768.

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 (2022): "Systemic Discrimination Among Large U.S. Employers," *The Quarterly Journal of Economics*, 137, 1963–2036.

——— (2024): "A Discrimination Report Card," *American Economic Review*, 114, 2472–2525.

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

Kohler-Hausmann, I. (2019): "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination," *Northwestern University Law Review*, 113, 1163–1227.

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

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," Working Paper.

Lundberg, S. J. and R. Startz (1983): "Private discrimination and social intervention in competitive labor market," *The American Economic Review*, 73, 340–347.

May, A., J. Wachs, and A. Hannák (2019): "Gender differences in participation and reward on Stack Overflow," *Empirical Software Engineering*, 24, 1997–2019.

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

McMillon, D. B. (2024): "What Makes Systemic Discrimination," Systemic?" Exposing the Amplifiers of Inequity," *arXiv preprint arXiv:2403.11028*.

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

Mocanu, T. (2023): "Designing Gender Equity: Evidence from Hiring Practices and Committees," IZA Discussion Paper 17480, Institute of Labor Economics.

National Archives and Records Administration (NARA) (2007): "Access to Archival Databases (AAD)," U.S. National Archives.

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.

Nunley, J. M., A. Pugh, N. Romero, and R. A. Seals (2015): "Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment," *The BE Journal of Economic Analysis & Policy*, 15, 1093–1125.

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

Oster, E. (2017): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37, 187–204.

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. (2008): "Structural Racism: Building Upon the Insights of John Calmore," *North Carolina Law Review*, 86, 791.

Rambachan, A., A. Coston, and E. Kennedy (2024): "Robust Design and Evaluation of Predictive Algorithms under Unobserved Confounding," arXiv preprint arXiv:2212.09844.

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

Rose, E. K. (2023): "A Constructivist Perspective on Empirical Discrimination Research," *Journal of Economic Literature*, 61, 906–923.

Roth, P. L., K. L. Purvis, and P. Bobko (2012): "A meta-analysis of gender group differences for measures of job performance in field studies," *Journal of Management*, 38, 719–739.

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

Ruebeck, H. (2024): "Causes and Consequences of Perceived Workplace Discrimination," Working Paper.

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: A Journal of Research*, 57, 509–514.

Sen, M. and O. Wasow (2016): "Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics," *Annual Review of Political Science*, 19, 499–522.

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

Strack, P. and K. H. Yang (forthcoming): "Privacy Preserving Signals," *Econometrica*.

Terrell, J., A. Kofink, J. Middleton, C. Rainear, E. Murphy-Hill, C. Parnin, and J. Stallings (2017): "Gender differences and bias in open source: Pull request acceptance of women versus men," *PeerJ Computer Science*, 3, e111.

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," in *Proceedings of the 26th International Conference on World Wide Web*.

Zivin, J. S. G. and G. Singer (2022): "Disparities in Pollution Capitalization Rates: The Role of Direct and Systemic Discrimination," Working Paper 30814, 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; Pager and Shepherd 2008; Feagin 2013; Allard and Small 2013). 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 2008). 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.[52] 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 et al. 2021). The literature sometimes refers to the former as "side-effect" discrimination and the latter as "past-in-present" discrimination (Feagin and Feagin 1978; Gynter 2003; 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.[53]

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

[52] Powell (2008) 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.

[53] 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 "color-blind" treatment because they were aware systemic discrimination could preserve the status quo.

as in *Griggs v. Duke Power Co. (1971)*.[54] 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.[55]

A systemic perspective is also found in the recent literature on algorithmic unfairness (e.g., Angwin et al. 2016; Hardt et al. 2016; Zafar et al. 2017; Gebru 2020; Berk et al. 2021; Kasy and Abebe 2021; Arnold et al. 2021; Buolamwini 2022). 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 and Mason 1998; Darity 2005; De Quidt et al. 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. 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. 2023; Hübert and Little 2023; Barron et al. 2024). These differences may stem from a lack of information or biased stereotypes (Fiske 1998; Bordalo et al. 2016, 2019; Coffman et al. 2021), which again lead to causal effects of a protected characteristic on evaluations and decision-making. As noted in the paper, our work also contributes to the rich literature on non-direct discrimination in economics considering "pre-market" forces and historical injustices in driving disparities (Neal and Johnson 1996; Glover et al. 2017; List 2004; Cook 2014; Hurst et al. 2024; Sarsons 2019). Our framework

[54] See also *Dothard v. Rawlinson (1977)* and *Cocks v. Queensland (1994)*.
[55] See, e.g., *Washington v. Davis (1976)* and *McClesky v. Kemp (1987)*.

puts structure on these disparities so they can be decomposed and measured as different systemic channels. In Section 3.3 we also discuss how the model of Coate and Loury (1993) captures a specific source of systemic discrimination in our framework.

A rich empirical literature in economics has used both experimental and observational data to identify the causal effect of group identity on treatment, holding other observables constant (e.g., Bertrand and Mullainathan 2004; Fang and Moro 2011; Bertrand and Duflo 2017). 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. 2022).[56] 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 et al. 2001; Grau and Vergara 2021) and quasi-experimental data (Arnold et al. 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 et al. (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.[57]

A growing literature in economics has examined the impact of historical 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. Derenoncourt

[56] While Kline et al. (2022) 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.

[57] 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. (2024) and Eli et al. (2023) 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 (2024) measure the systemic components of disparities that stem from racially oppressive institutions—slavery and Jim Crow laws. Baron et al. (2024) 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 (2022) 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 present a theoretical example and an experiment to illustrate how group-based differences in the precision of productivity signals can lead to both direct and systemic discrimination in a screening action. The former channel is through accurate statistical discrimination: the groups face different effective thresholds for the same signal realizations because of the difference in signal precision. The latter systemic channel comes from the difference in the signal distribution, accounting for the difference in thresholds. For example, if an aptitude test is designed by a dominant group it may provide more accurate information about members of that group than for a minority group; alternatively, a medical diagnostic test may only be trialed on the majority group and is thus more predictive for this group. Such disparities in screening accuracy correspond to a type of systemic discrimination: even if individuals from different groups receive the same treatment conditional on the same test result, if the system neglects developing accurate methods to screen minority groups, these groups will face systemic discrimination.

Both the theoretical example and the experiment show that canonical statistical discrimination models may not capture the full extent of (total) discrimination stemming from differences in the signaling technology. They also show how discrimination due to differences

in the signaling technology manifests in fundamentally different ways than discrimination due to differences in the prior distribution of productivity (i.e., the other source of classic statistical discrimination). When the qualification is set to current productivity, $Y_i^0 = Y_i^*$, the former can lead to both direct and systemic forms of discrimination in the current decision, while the latter only leads to direct discrimination. Finally, we show how systemic discrimination from disparities in the informativeness of signals is likely to be heterogeneous across worker productivity levels: more productive workers tend to face more systemic discrimination than less productive workers.

B.1.1 Theoretical Example

Suppose worker productivity is distributed identically within groups, $Y_i^* \sim N(0,1)$, but the manager's signal $S_i = Y_i^* + \varepsilon_i$ has a group-specific precision: $\varepsilon_i \sim N(0, 1/\eta_g)$ when $G_i = g$, with more precise signals for group w, $\eta_w > \eta_b > 0$. The distribution of S_i for a group-g worker with productivity y is $N(y, 1/\eta_g)$ and the posterior expected productivity for a worker from group g who generates signal realization s is $s\eta_g/(1+\eta_g)$. This example sets productivity as the qualification, $Y_i^0 = Y_i^*$.

Suppose the manager hires all workers whose posterior expected productivity is at or above some threshold $t \in \mathbb{R}$: $A(g,s) = \mathbb{1}\{s\eta_g/(1+\eta_g) \geq t\}$. The manager thus hires group-g workers with signal realizations $S_i \geq t(1+\eta_g)/\eta_g$. Group-$b$ workers face a higher signal threshold, since $(1+\eta_b)/\eta_b > (1+\eta_w)/\eta_w$. Therefore, there is direct discrimination against group b stemming from the higher cutoff arising from their less precise productivity signal. Specifically, group-w workers with $S_i \in (t\frac{1+\eta_w}{\eta_w}, t\frac{1+\eta_b}{\eta_b}]$ are hired but group-b workers with signals in this range are not (hiring of workers with other signals does not depend on group).

Even without the direct discrimination in signal thresholds, however, the difference in signal precision causes equally-productive workers to be hired at different rates depending on their group. For a given $y \in \mathcal{Y}$ and $g \in \{b, w\}$, systemic discrimination is captured by

$$\begin{aligned}
&E[A(g,S_i)|Y_i^* = y, G_i = w] - E[A(g,S_i)|Y_i^* = y, G_i = b] \\
&= Pr\left(S_i \geq t(1+\eta_g)/\eta_g | Y_i^* = y, G_i = w\right) - Pr\left(S_i \geq t(1+\eta_g)/\eta_g | Y_i^* = y, G_i = b\right) \\
&= \Phi\left(\eta_b\left(t(1+\eta_g)/\eta_g - y\right)\right) - \Phi\left(\eta_w\left(t(1+\eta_g)/\eta_g - y\right)\right),
\end{aligned}$$

where $\Phi(\cdot)$ gives the standard normal distribution.

Since $\eta_b \neq \eta_w$, this expression is non-zero unless $y = t\frac{1+\eta_g}{\eta_g}$. Therefore, there is systemic discrimination almost everywhere in the productivity distribution, stemming from the differential probabilities of the signal being above a given cutoff for equally productive group-w versus group-b workers.

Systemic discrimination in this screening action is heterogeneous across worker produc-

tivity levels. With $\eta_w > \eta_b > 0$, the systemic discrimination hurts group-b workers at high levels of productivity (where $y > t\frac{1+\eta_g}{\eta_g}$) and *favors* group-$b$ workers at low levels of productivity (where $y < t\frac{1+\eta_g}{\eta_g}$) since $\Phi(\cdot)$ is strictly increasing. Intuitively, having a higher signal variance makes low-productivity group-b workers more likely to have a signal above the effective threshold by chance, while high-productivity group-b workers are more likely to generate a signal below the threshold by chance.

The average level of systemic discrimination across workers depends on which of these two productivity groups is larger. In a "cherry-picking" market with $t > 0$, such that a minority of workers are hired in each group (i.e., $Pr\left(S_i \geq t\frac{1+\eta_g}{\eta_g} \mid G_i = g\right) < 0.5$), the systemic discrimination favors group-b overall. This is because there are fewer high-productivity group-b workers hurt by the higher signal variance than low-productivity group-b workers helped by it. Conversely, in a "lemon-dropping" market ($t < 0$) with most workers being hired, the systemic discrimination hurts group-b workers overall.

This theoretical example highlights how examining screening bias with only a direct measure of discrimination may miss an important component of how differential signal precision impacts total discrimination.

Similar to the case of statistical direct discrimination (e.g., Fang and Moro 2011), differential signal precision can be heterogeneous across qualification levels. 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 increasing the chances of the generated signal exceeding the hiring threshold. In contrast, in a medical diagnostic context, all patients benefit from a more accurate signal when it leads to more accurate diagnoses, regardless of their health status.

B.1.2 Experimental Setup

Workers: 100 participants were randomly assigned to the role of Worker. Each Worker completed two sets of tasks (A and B) and provided basic demographic information including self-reported group identity G_i (either male m or female f). Each task consisted of a test of the Worker's basic math, business, and history knowledge, with 10 randomly selected questions from these subjects. A Worker's performance on each task was defined as the number of questions she answered correctly. We restrict attention to Workers with a task-A performance in $\mathcal{S}^R = \{2, 3, 4, 5, 6\}$ in order to ensure enough data for each gender.

There were no significant gender differences in Worker performance on either task. On average, Workers completed 3.57 questions correctly on task A and 3.53 questions correctly on task B. Regressing overall performance (the sum of performance on both tasks) on a male Worker indicator yields an insignificant coefficient of -0.13 ($p = 0.84$). The gender coefficient

is similarly insignificant when we regress performance on task A (0.21; $p = 0.63$) and task B (-0.34; $p = 0.35$) separately. Performance on task B was predictive of performance on task A. Regressing the latter on the former yields a coefficient of 0.36 ($p < 0.01$). Furthermore, there were no significant gender differences in this relationship: regressing task-A performance on task-B performance, gender, and their interaction yields an insignificant interaction coefficient of 0.15 ($p = 0.58$).

Recruiters: 199 participants were randomly assigned to the role of Recruiter. Each Recruiter was 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, beginning after the fifth correct answer. The Recruiter's action rule is thus $A_i^R \in \{0, 1\}$.

Hiring Managers: 501 participants were randomly assigned to the role of Hiring Manager. Each 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.

B.1.3 Results

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. 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%.[58] Given the lack of gender-based performance differences, as reported in Appendix B.1.2, this disparity in hiring rates is not consistent with accurate statistical discrimination. Therefore, Recruiter direct discrimination 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 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 ($p = 0.43$).[59] Among Workers with high (above-median) qualification levels, male Workers were hired at a significant 23 percentage point higher rate ($p < 0.01$).

Figure 7 illustrates the reason for this heterogeneity in total discrimination. Similar to Figure 6, the scatter plot shows the average Hiring Manager actions conditional on the signal (or lack thereof) and the Worker's gender. 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 (at or above the median score) 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 from 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.

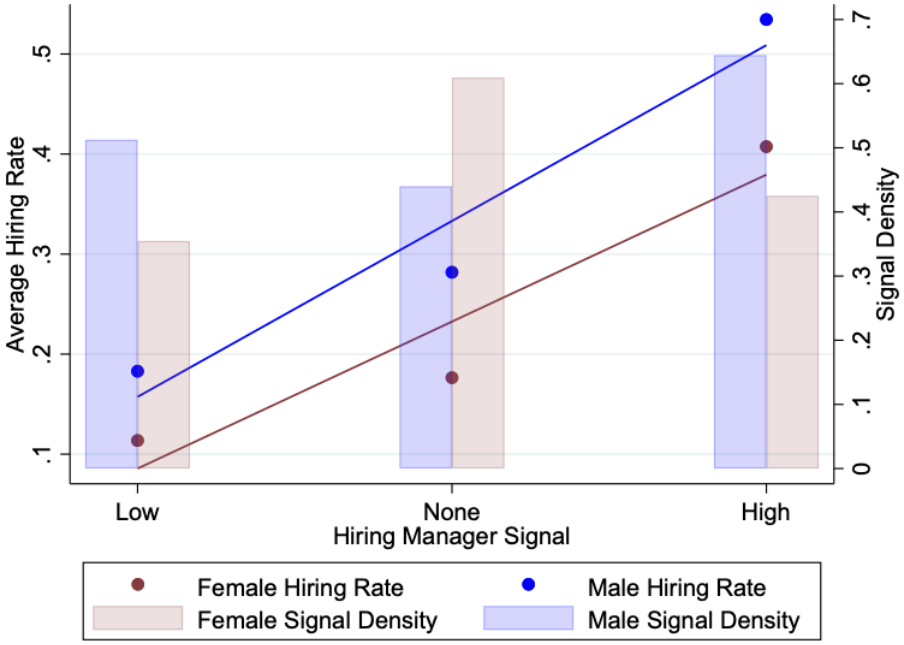
We quantify total, direct, and systemic discrimination in Hiring Manager actions using the decompositions in Section 3.2. We estimate Hiring Manager total discrimination $\Delta(y^0)$ by comparing male and female hiring rates based on task-A performance. We then estimate the Hiring Manager average direct discrimination $\overline{\tau}(w, y^0)$ faced by male Workers with a given task-A performance by averaging the gender disparities across each Hiring Manager signal realization according to the distribution each task-A performance induces over this signal. Subtracting this value from the estimate of total discrimination yields an estimate of the measure of systemic discrimination.[60] 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.

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

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

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

FIGURE 7. Screening: Hiring Manager Hiring Rate and Signals



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

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 24%. Our decomposition shows this is driven by a combination of significant direct (15%) and systemic discrimination (9%). In contrast, total discrimination among workers with a low qualification is small and insignificant (3%), despite significant direct discrimination (7%). The reason is a small degree of negative systemic discrimination among less qualified Workers (-4%). Consistent with the model in Appendix B.1.1, the gap in systemic discrimination across qualification levels is significant ($p = 0.04$).

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.
*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

C Experimental Details

C.1 Constructive IA

C.1.1 Scraping Job Listings

For the IA Study

We scraped job listings from corporate career websites of five major automotive firms. For each company, we scraped open entry-level job listings for the 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 was 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 "experi-

ence/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 stores 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, except 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 labeled as being in a specific city; for companies that allowed searching for jobs within certain radius of a location, we chose jobs within a 5-mile radius of the city (the smallest radius). A research assistant similarly monitored the process.

C.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 within $\pm$ 0.29 degrees and $\pm$ 0.37 degrees of the job location's latitude and longitude, respectively. Such degrees approximately correspond to 20 miles in distance. Further, we assigned an ordinal variable to the addresses: a value of 0 if the address is in the same city as the job location, 1 if the address is in the same state but a different city, and 2 if the address is in a different state. 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 used 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 had 200 rows to randomize from, given the scarcity of residences 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 relabeled 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 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 variable (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 a uniform distribution $[0.5, 2]$ 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 graduation date, 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 set as the high school graduation date plus the number of unemployment days before prior employment, and the employment end date was set as 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." 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, for a zip code of 60637 the range would be $[60537, 60737]$. Similar to the residential addresses, we then assigned an ordinal variable to each high school depending on whether it 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 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 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 appli-

cant was randomly chosen from four potential formats: first-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 of 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 a phone bank firm, and were set up such that all calls and texts were redirected to a single number that was monitored. 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 a randomly selected format: either "XXX XXX XXXX" or "(XXX) XXX-XXXX".

Volunteer Experience

We included one volunteer experience for every fictitious applicant. 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.

Template

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

C.1.3 Filing Online Applications

Birthday and Age

Each fictitious applicant had a high school graduation age between 17 and 19. If the

applicant had no prior employment, we assumed their 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 their work experience in years. To determine the applicant's birthday, we first calculated their birth year by subtracting from the year when the experiment was run their current age (last-year age plus 1). Then, we added a randomly selected number of days from uniform $[0, 364]$ to the first day of their birth year to arrive at their exact birth date. This method ensured all applicants were over 18 at the time of application, preventing age limits from confounding the results.

SSN

In some cases, SSNs were required in the online applications. We assigned each fictitious applicant an SSN from the publicly available database of SSNs belonging to people deceased before 2007 (on National Archives and Records Administration (NARA) 2007). We manually selected different SSN files based on surname initials and varied starting digits 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 at different times across three days. For example, if we sent the first application on Monday morning, each of the remaining three applications was sent during randomly selected blocks from the following eight: Monday afternoon, Monday evening, Tuesday morn-

ing, 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 prior employment (if any) was at 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 used another job for the unsent conditions, such that all four treatment arms were applied across the two job listings.

C.1.4 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 the dedicated phone line, we identified phone numbers associated with the companies we applied to by examining their displayed names (e.g., some numbers will show the name of the company) and their voicemail. This allowed us to create a dictionary of verified phone numbers and their associated companies. Second, we filtered phone calls from these verified numbers, and matched them to our fictitious applicants based on the recipient's (applicant's) phone number. In the cases where the same company called the same applicant multiple times, we only recorded the first call from this company. This resulted in a dataset of applicant-company pairs where at least one call was made. Lastly, we mapped all of the applicant-company pairs to the original dataset of applicants, associating the first-time calls with the treatments. We then summarized the results across the treatment arms to arrive at the final estimates of first-time calls for each treatment arm.

C.1.5 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.

C.1.6 Callback Converstion Rate

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

We first asked them for 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 to provide 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 was randomized within each block. Afterwards, the surveyed managers answered the demographic questions and one open-text question for them to leave comments.

C.2 Experimental IA

C.2.1 Preparing Letter Content

Focusing on STEM majors, we 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 based on location and primary SIC code (indicating industry) to only preserve companies whose industries are related to these two majors, and which are located in cities near 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 sorted based on any variables, selecting 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 gap between employments 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 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 identical education and job experience. The two pairs took up the two majors, corresponding universities, and related experiences, respectively.

C.2.2 Generating Example Letters

With this information, we prompted ChatGPT to write recommendation letters by asking for a 300-word letter using all the education and prior experience information for each candidate. All such information was provided in bullet-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 exhibited similar levels of bias in lexical content across 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.

C.2.3 Generating Multiple Letters

Once we generated the initial four letters, we changed the ChatGPT prompt to request additional letters, similar in terms of sentiment and type of words, for the same candidates and the same positions outlined in the previous subsection. In this manner, we created 25 letters for each candidate. These correspond to the Endogenous-m and Endogenous-f 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-f letters. We used the same process to create the Exogenous-m letters.