(Mis)anticipated Discrimination

Ethan O'Leary*

Job Market Paper.

Last Updated: September 22, 2026.
Click here for latest version.

I propose and test a novel mechanism through which discrimination persists: the misperception of statistical discrimination as taste-based discrimination. I first derive a model of self-confirming discrimination in a job-application setting, in which misperceiving statistical discrimination as taste-based is a key channel: correctly attributed statistical discrimination admits both discriminatory and non-discriminatory equilibria, so inequality reflects a coordination failure, while misattribution to taste-based origins yields only discriminatory equilibria and lowers the perceived return to re-coordination. In a pre-registered laboratory experiment, subjects who experience discrimination from an ambiguous origin overestimate the probability that this inequality is caused by taste-based discriminators. This misattribution is associated with a higher probability that application strategies sustain statistical discrimination: a 1 percentage point increase in believed probability that discrimination is taste-based corresponds to a 0.35 (0.24) percentage point decrease (increase) in the probability that a subject in the discriminated (favoured) group applies to a job, perpetuating statistical discrimination. Moreover, the more likely subjects perceive the employer to be a taste-based discriminator, the more unfair they perceive the labour market and the lower their demand for a temporary affirmative action policy aimed at re-coordinating toward a less discriminatory outcome. This demonstrates that the persistence of discrimination is not only driven by the underlying source, but also the perception of such.

JEL: J7, D91, J2, C92
Keywords: Discrimination, Misspecified Beliefs, Affirmative Action, Labour Supply

*FAIR The Choice Lab & Department of Economics, Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway. Email: ethan.oleary@nhh.no.
I am indebted to Erik Ø. Sørensen and Heidi Thysen for their invaluable guidance in this project. Moreover, I would like to thank Kai Barron, Kfir Eliaz, Nickolas Gagnon, Michael Hilweg-Waldeck, Anna Hochleitner, Dorothea Kübler, Jonas Pilgaard Kaiser, Fidel Petros, Sebastian Schweighofer-Kodritsch, Morten Nyborg Støstad and Georg Weizsäcker for helpful comments. Additionally, I would like to thank José Guinot Saporta for his excellent assistance and flexibility in running the laboratory experiment. I would also like to acknowledge participants of The Choice Lab Coffee Meeting (FAIR) for additional input. Pre-registration for the experiment can be found at https://www.socialscienceregistry.org/trials/17144. This project has been supported financially by the Professor Wilhelm Keilhaus Minnefond and Norwegian School of Economics' Centre for Ethics and Economics Grant. Ethical approval was received from NHH Institutional Review Board protocol number NHH-IRB-2025-131.

1 Introduction

Minorities continue to be under-represented across the labour market and discrimination is a leading candidate explanation for this (Bertrand and Duflo, 2017). In studying this phenomenon, researchers have distinguished between two sources of unequal treatment: taste-based discrimination is the animus towards another group potentially manifested through a psychological cost of intergroup interactions (Becker, 1957), and statistical discrimination, driven by unobservable qualities, is the inference of individual characteristics from group level beliefs (Phelps, 1972; Arrow, 1974).[1] The literature studying the dynamics and stability of discrimination has focused on understanding the underlying source and the trajectory of each source over time (Pager, 2003; Eberhardt et al., 2004; Goldin, 2014; Bertrand and Mullainathan, 2004; Blau and Kahn, 2017; Bohren et al., 2025). From this endeavour, Bohren et al. (2019) show that the source of discrimination is detrimental to the longevity of unequal treatment: if two groups are ex-ante equivalent except for some uninformative identifying characteristic, then statistical discrimination, driven by information asymmetry, should dissipate as information is generated over the career cycle. On the other hand, taste-based discrimination may survive as long as preferences do. In this paper, I demonstrate theoretically and experimentally that the persistence of discrimination relies on both the true underlying source and how that source is perceived: that statistical discrimination may persist if it is incorrectly anticipated to be taste-based discrimination

The behavioural responses of workers to the anticipation of discrimination, particularly in human capital formation and labour supply decisions, have received much recent attention (Angeli et al., 2023; Boring et al., 2025; Gagnon et al., 2025; Lepage et al., 2025; Ruebeck, 2024; Ugalde, 2022). Far from passive reactions, these adjustments shape lifetime earnings, career trajectories, and social mobility for entire groups. Despite being rational responses to beliefs, the behaviours can create a self-fulfilling dynamic: two ex-ante identical groups of agents who expect differential treatment and respond by taking separating actions will appear ex-post different. This divergence can force an information-constrained, profit-seeking principal to make group-based distinctions, reinforcing the very beliefs that produced it (Arrow, 1974; Coate and Loury, 1993; Glover et al., 2017; Dianat et al., 2022). Alternatively, expectations of equal treatment provide no incentive for groups to take separating actions and so the unbiased principal discards group identity in their evaluation decision. These self-confirming equilibria reveal discrimination to be a result of a market coordination error which arises solely because it is believed to do so. This paper shows that misperceiving this mechanism may lead to persistent discrimination.

I first derive a novel theoretical model which embeds the self-confirming discrimination concept of Coate and Loury (1993) within the job search framework of Lucas (1978). Workers belong to one of two groups and draw a productivity level from a common, group-independent distribution. In each period, workers find an employer offering a fixed number of vacancies at an exogenous wage, and must choose between submitting a costly application or pursuing an outside entrepreneurial option that yields a return equal to own productivity.[2] Each applicant is eligible for unemployment insurance if they are unsuccessful. An employer makes hiring decisions based on the observable information of each candidate: their group identity and a noisy unbiased signal of their productivity. Following Lucas (1978), best-response strategies take cut-off forms: workers apply if and only if their productivity lies below some threshold and employers, anticipating this,

[1]Often one may decompose statistical discrimination according to how the necessary informational frictions arise. Informational frictions may arise because the informativeness of signals is group-contingent (Phelps, 1972) or because ex-ante identical groups coordinate on different strategies which lead to ex-post differences that a potential discriminator may believe in to make evaluations (Arrow, 1974).

[2]The terminology of an entrepreneurial outside option is used to remain close to the setting of Lucas (1978); one may interpret it more broadly, such as remaining in job search.

hire applicants with signals at least as great as some critical value. I define discrimination as the case in which the hiring thresholds differ according to groups.

When workers anticipate no discrimination, their perceived hiring probability is independent of group identity, and applications and hiring strategies correspond to a non-discriminatory equilibrium. When workers instead anticipate discrimination, the probability of a worker with some productivity level being hired depends on their group identity and so separating application strategies are adopted, generating self-confirming discrimination. In this equilibrium, the disadvantaged group is disproportionately pushed into entrepreneurship and earns lower aggregate pay than the favoured group. If workers correctly attribute this discrimination to its statistical origin, they would recognise the underlying coordination failure attributing to the outcome: were all groups to coordinate on identical application cut-offs, discrimination would vanish. Whether workers actually reach this attribution, however, is an open empirical question.

I address this with a representative survey, detailed in Appendix A, which asks how individuals perceive the prevalence of discriminatory attitudes in the population. The survey reveals that although prejudicial preferences appear to have declined over the past century, respondents today systematically overestimate the share of others who hold discriminatory attitudes towards minority groups. This pattern suggests that individuals may overestimate the probability of encountering a taste-based discriminator, or misattribute the source of realised discrimination to preferences rather than statistics. Motivated by this evidence, I allow workers in the model to misperceive employers' preferences and beliefs, leading them to *mis*anticipate taste-based discrimination. Under this misperception, workers expect discrimination to be inevitable regardless of their application strategy: a non-discriminatory equilibrium is not perceived as attainable.

As the discriminatory equilibria under correct and incorrect attribution are observationally equivalent, arriving at this outcome represents an identification challenge for the worker. Observing discrimination alone reveals nothing about its source; workers can only learn the true cause by experimenting with different application strategies and observing the resulting hiring outcomes. As equilibrium coordination is grounded in beliefs of others' behaviours, the model derives a testable hypothesis: that the more likely agents perceive discrimination to be taste-based, the more likely they are to coordinate on a discriminatory equilibrium. If true, misattribution becomes a trap: discrimination persists, and agents do not learn the true source.

I test the model's predictions: that if individuals misattribute the source of discrimination they are more likely to play application strategies that sustain statistical discrimination, using a laboratory experiment. Sixteen participants, equally split between two artificial identities, formed experimental societies to play a repeated job application game modelled on the theoretical setup above. In each round, participants drew a productivity value and decided whether to apply for a job. They were informed that, at the beginning of the experiment, one of two employer types would be randomly chosen: a profit-maximising statistical discriminator or a taste-based discriminator, however participants were unaware of the pre-registered randomisation rule and that the random draw under this rule resulted in all societies facing the former type. After observing the identity of those hired in their society in the current round, participants had to report their belief about the employer's type.

I find that subjects who experience discrimination consistent with the preferences of the taste-based discriminator overestimate the probability that the employer is a taste-based discriminator. Those who observe at least two thirds of hiring decisions going to the favoured group believe that there is a 57.0% probability that the employer is a taste-based discriminator, significantly above the Bayesian benchmark of 50%. Crucially, at a given history of discriminatory hiring, a 1 percentage point increase in the subjective probability that the employer is a taste-based discriminator lowers the estimated probability that a candidate from the discriminated group applies to the job by 0.35 percentage points and increases that of the favoured group by 0.24 percentage points. This effect is particularly pronounced among subjects with intermediate productivity values: those predicted to be on the margin of application. Therefore, the more likely subjects perceived the employer to be a taste-based discriminator, the more separated predicted application strategies and the more separated the average productivity of applicants from each group. Misattribution, in other words, is associated with a higher probability that application strategies sustain discrimination.

Misattribution may also sustain discrimination indirectly through two supplementary channels. First, at a given level of inequality, the more likely a prejudiced individual perceives the discrimination to come from taste, the more unfair they perceive the labour market. This raises the possibility of backlash effects, whereby individuals withdraw their labour supply from markets in which they perceive the playing field to be unjust (Charness and Levine, 2000; Offerman, 2002; Abeler et al., 2010; Gagnon et al., 2025). Second, misattribution reduces demand for, and the perceived efficacy of, temporary affirmative action (Coate and Loury, 1993). Such a policy would help re-coordinate strategies toward the non-discriminatory equilibrium, generating long-lasting equality gains even after the policy is lifted (Gupta, 2024). Yet the misperception of the true source of discrimination draws a veil over this possibility, potentially leading to its suboptimal provision.

This paper contributes to the discourse on the emergence, persistence, and consequences of discrimination by introducing the novel concept of discrimination misattribution, or the *misanticipation* of its source. In doing so, I speak to a nascent but growing literature on the effects of anticipated discrimination on labour supply decisions (Card et al., 2012; Bracha et al., 2015; Glover et al., 2017; Breza et al., 2018; Flory et al., 2018; Leibbrandt and List, 2018; Samek, 2019; Gagnon et al., 2025).[3] Close to my paper are the recent experiments of Angeli et al. (2023); Ruebeck (2024) and Boring et al. (2025) who find that beliefs of discrimination shape the composition of the applicant pool. I add to this literature by theoretically and experimentally modelling how labour demand responds to these supply-side distortions, closing the loop between anticipated discrimination and its equilibrium consequences.

I also extend the theoretical and applied literature on self-confirming discrimination. In their canonical model, Coate and Loury (1993) show that individuals from groups expecting discrimination underinvest in human capital formation because they correctly perceive lower returns to such acquisition leading to equilibrium discrimination. Their multiple-equilibria framework demonstrates how statistical discrimination can be understood as a coordination failure whenever a feasible non-discriminatory equilibrium also exists. Subsequent extensions have incorporated intergroup interactions (Moro and Norman, 2003), private and public sectors (Fang and Norman, 2006), and employer investment in training (Craig and Fryer, 2017), while self-confirming discrimination has been documented empirically both in the field (Glover et al., 2017) and in the laboratory (Dianat et al., 2022).[4] I extend this framework further by embedding it within the job search domain of Lucas (1978) and Poschke (2013), offering a novel explanation for observed application gaps in the labour market and demonstrating potential equilibrium effects of such (Fluchtmann et al., 2024).

I argue that one should be concerned not only with whether individuals anticipate discrimination, but also with which source of discrimination they anticipate. I show empirically that incorrect

[3]Other responses to beliefs of discrimination have been studied such as human capital investment (Ugalde, 2022; Lepage et al., 2025), over-investment in productivity signals (Lang and Manove, 2011; Agüero et al., 2023; Exley et al., 2024), and even paying to mask one's identity (Charness et al., 2020; Alston, 2025; Aksoy et al., 2023) or masquerading as another (Kudashvili and Lergetporer, 2022; Gerdeman, 2017).

[4]My work also builds on the results of Filippin (2009) who demonstrates how anticipating discrimination that is never to arise can induce bad equilibria in an effort-provision and promotion setting. This represents an inequality trap as the true share of discriminators declines. In contrast, discrimination in my model is based on the Bayesian inference of an unobservable ability.

beliefs of the source of discrimination lead to a stickier self-confirming equilibrium.[5] In doing so, this paper relates to a large literature on behavioural biases of agents when observing sampling from defined distributions. My results demonstrate that individuals do not account for the Bayesian inference from truncated distributions speaking to the literatures on the so-called flat tail bias (Peterson et al., 1968; Wheeler and Beach, 1968; Teigen, 1974; Benjamin et al., 2017), inattention to true probability distributions (Zhang and Maloney, 2012; Steiner and Stewart, 2016), and behavioural attenuation to complexity (Frydman and Jin, 2022; Enke and Graeber, 2023; Enke et al., 2024, 2025; Gabaix and Graeber, 2024).[6] I show that individuals tend to take the path of least mental resistance when updating their beliefs in response to uninformative signals, linking to the literatures on misperceptions of others' strategies (Eyster and Rabin, 2005), representativeness (Rabin and Vayanos, 2010; Gennaioli and Shleifer, 2010; Bordalo et al., 2016) and contingent thinking (Martínez-Marquina et al., 2019). In particular, I relate closely to the work of Haggag et al. (2019); Enke (2020); Graeber (2023) and Barron et al. (2024) who demonstrate that misattribution of the data-generating process can lead to suboptimal outcomes.

The paper continues in the following manner. Section 2 outlines my theoretical model of self-confirming discrimination and demonstrates that misattribution of discrimination sustains discriminatory equilibria. In section 3, I detail a laboratory experiment in which I test the main hypotheses derived from my theoretical model. I present the results of this experiment in section 4, in which I show that subjects systematically misattribute the source of discrimination and this has direct and indirect consequences on the persistence of inequality. In section 5, I discuss the results in light of the literature and attest to policy implications regarding the re-coordination of the labour market. I also discuss evidence that the underlying misperception driving misattribution may be linked to the complexity in the attribution process. Finally, in section 6, I conclude.

2 Theoretical model

The theoretical model demonstrates that the set of equilibria depends on the attribution of discrimination source. When agents correctly identify the employer as a statistical discriminator, multiple equilibria may exist: a non-discriminatory equilibrium is always available, and arriving at a discriminatory outcome constitutes a coordination error that internal adjustment or external policies such as affirmative action can resolve. When the employer is misperceived as a taste-based discriminator, the non-discriminatory equilibrium is not perceived as attainable. In this case, a discriminatory outcome is perceived as an inevitability driven by the employer's preferences rather

[5] There is much recent discourse on how discrimination may be sustained by irrational agents, however mainly from the demand side of the market (see Onuchic (2022) for a review). Employers may develop biased priors due to overweighting group differences (Bordalo et al., 2016); through misinterpreting the sampling procedure (Lepage, 2024; Bohren et al., 2025); or through violations to Bayesian inference (Sarsons, 2017; Eyting, 2022; Rackstraw, 2022; Esponda et al., 2023; Campos-Mercade and Mengel, 2024; Ruzzier and Woo, 2023). In particular, a number of papers have proposed that employers may discriminate based on inaccurate beliefs (Chauvin, 2018; Hübert and Little, 2023; Ruzzier and Woo, 2023; Bohren et al., 2025), a matter which Bohren et al. (2019) argue is dynamically unstable, as employers should become aware of such inaccuracies when hiring occurs under informational frictions.

[6] An alternative interpretation of my framework is that individuals have a misspecified model of the signal generating process that they observe: they have an incorrect belief of the distribution of signals available (see Bohren and Hauser (2025) for a review). Misattribution can therefore be interpreted as a misspecified model of the discrimination in which individuals learn inefficiently of the true state of the world (Bohren and Hauser, 2021; Fudenberg et al., 2021; Frick et al., 2023; Esponda et al., 2021). These papers operate using a Berk-Nash equilibrium framework in which agents with misspecified models converge their beliefs to a subjective truth. I do not go insofar as to formalise this equilibrium concept as the subjective model is equivalent to beliefs of discrete principal types such that my equilibrium definition is less restrictive.

than a correctable coordination failure.

I build on the model on the canonical framework of self-confirming discrimination developed by Coate and Loury (1993). In this seminal job-assignment model, employers observe workers' group identity but not their productivity, and assign tasks based on a belief of the correlation between the two. Workers decide whether to invest in a costly productivity-enhancing qualification before assignment, drawing an idiosyncratic investment cost from a known, group-independent distribution. If an employer harbours a negative stereotype against group X, believing the average productivity of X workers is lower than that of other groups, the financial incentive to assign X workers to the more demanding, better-paid task falls, which in turn lowers X workers' own incentive to invest. Because employers' beliefs are themselves shaped by, and, in equilibrium, must remain consistent with, workers' investment behaviour, this beliefs of discrimination can become self-confirming: even though groups are ex-ante identical, negative stereotypes can persist as a stable outcome. Critically, alongside this discriminatory equilibrium there also exists a non-discriminatory equilibrium in which no stereotypes are held and both groups invest at equal rates.

The model presented in this section departs from Coate and Loury (1993) by inverting the action and type sets. In their framework, labour supply is fixed and workers choose an unobservable costly investment in skill acquisition. In this paper, I fix the distribution of workers' human capital and endogenise their labour supply decisions instead. This is motivated by a recent applied literature which has focused on the effects of beliefs of discrimination on labour supply decisions noting that gaps in labour supply emerge after human capital investment decisions. At this point, beliefs of discrimination in the labour market may arise (Angeli et al., 2023; Ruebeck, 2024; Boring et al., 2025). This anticipatory behaviour may then explain patterns such as occupational sorting (Goldin, 2014) and application gaps (Fluchtmann et al., 2024; Bandiera et al., 2025).

Under this structure, workers are endowed with an exogenous productivity value from a known distribution and choose between applying for a fixed-wage job or taking their productivity value as an entrepreneurial outside option. This structure is based on the job search model of Lucas (1978) in which there is negative selection into job applications.[7]

I first derive the case of correct attribution: where statistical discrimination is correctly anticipated. In this instance, multiple equilibria exist and discrimination is not the unique stable outcome. Using a representative survey of a US-based population, I find that respondents overestimate the share of others harbouring prejudicial preferences against minorities and thus, it may be unreasonable to assume perfect attribution of discrimination source.[8] In subsection 2c, I relax the assumption that workers are aware that the employer is truly not a taste-based discriminator and develop a model of misattribution.

2a Setting

Workers. There is a unit mass of risk-neutral workers denoted by $i \in [0, 1]$, each of whom lives for one period. Workers are given an exogenous type which constitutes an observable group

[7] This is a key assumption of the present model which somewhat departs from standard economic models in which positive selection into work is the primal feature. The present structure can be motivated by considering the large number of occupations in which one may opt to work for an employer with a fixed wage, perhaps minimum wage or jobs in which individual bargaining is limited, or choose to operate as a sole trader. Examples include taxi drivers, janitors, tradespeople, florists, hair dressers, personal trainers, among others.

[8] The survey asked participants whether they approved of interracial marriage and if they would vote for a well-qualified presidential candidate from their party that was female. Next, participants were incentivised to state their belief of the share of respondents who negatively responded to these questions. The sample was collected on Prolific in 2025 and was representative according to age, gender, ethnicity and political affiliation. The survey is detailed in Appendix A.

identity $g_i \in \{A, B\}$ and an unobservable level of productivity $q_i \in \mathbb{R}$. Types are distributed according to a commonly known process: the share of A group identities is n_A and productivity is distributed independently and identically according to the function $F(q)$ and density $f(q)$. I assume that $F(q)$ is a normal distribution with mean $\mu > 0$ and concentration $\tau \in [0, 1]$.

Employer. A risk-neutral employer can hire up to an exogenous mass ϕ of workers to conduct a job which pays a fixed wage $\omega > 0$. The employer is given an exogenous and unobservable type which constitutes two attributes. First, they harbour preferences for hiring a worker of group A over one of group B. These preferences are denoted by $m = M^A - M^B$ where $M^g > 0$ is the marginal utility of hiring an individual from group g. Second, the employer has a level of sophistication $\psi \in \{0, 1\}$, which refers to their ability to anticipate the strategies of workers.

I assume that the employer is always unbiased, $m = 0$, and sophisticated, $\psi = 1$. Workers have beliefs of the employer's type $\hat{m}$ and $\hat{\psi}$, which define the attribution of discrimination source. I will later consider the case in which workers incorrectly believe the employer to be biased $\hat{m} \neq 0$.

Actions and payoffs. Each worker i must choose whether to apply for a job with the employer. If they do not apply, then they take up the entrepreneurial outside option which derives entrepreneurial rents equal to their productivity q_i.[9] The application is financially costless but time-consuming meaning that workers cannot apply and simultaneously engage in entrepreneurship.[10] If their application is successful then they are hired and earn a wage ω; if they are unsuccessful, their payoff is 0. One can interpret this as an unemployment insurance contingent on sending applications in the labour market.

For each worker i who applies, the employer simultaneously observes their group identities and a signal of each of their productivities, and decides which applicants to hire. The signal is noisy and unbiased, denoted by $s_i = q_i + \epsilon_i$ where ϵ_i is an independently and identically normally distributed noise with mean 0 and variance $1/\eta$. The employer is protected against individual productivity risk, so the payoff from hiring any worker is bounded below by 0.

Strategies. Given that the payoff from hiring worker i is strictly increasing in their productivity q_i and that the signal of productivity s_i is unbiased with full support, the expected return to hiring an applicant is strictly increasing in the signal that they portray. Given the labour demand constraint ϕ, the employer's best response strategy is to hire all applicants from group g who portray a signal at least as great as some threshold $\underline{s}_g$ for each group $g \in \{A, B\}$.

Denote the worker's belief of the employer's best response strategy for some group g to be $\hat{\underline{s}}_g$, then a worker will apply to the job if and only if their productivity is not larger than their expected earnings at the firm,

$$q_i \leq \omega \cdot Pr(s \geq \hat{\underline{s}}_g | q_i). \tag{1}$$

To simplify proceedings such that strategies are in line with those of Lucas (1978), I make the following assumption to ensure that there is a single level of productivity such that (1) holds with equality.

Assumption 1: The pecuniary returns to receiving a job are bounded by the informativeness of

[9] This a simplification of the set-up by Lucas (1978) who assumes the entrepreneurial productivity to monotonically increasing in productivity. Thus, the underlying assumptions are that each individual is perfectly aware of their own productivity and that either their entrepreneurship can contract on their own effort and output, or the expectation of entrepreneurial output is equal to their productivity.

[10] To support this assumption, one may consider the case in which an individual has a time endowment which may be spent in application preparation or setting up the entrepreneurial endeavour. As long as these costs are independent of group identity, the results hold.

the signal,

$$\omega \leq \sqrt{\frac{2\pi}{\eta}}. \tag{2}$$

Intuitively, this bounds the wage premium from employment relative to the informativeness of the productivity signal, preventing workers from always preferring to apply regardless of their outside option. With this assumption, it is possible to show that for any belief of employer strategy, there is a unique level of productivity for which a worker is indifferent between applying for the job and taking the entrepreneurial outside option. Those with productivity greater than this value will take the outside option and those with lower productivities will apply to the job. This represents a unique cut-off strategy for the workers.

Lemma 1: Existence of a unique cut-off strategy. At any belief $\hat{\underline{s}}_g \in \mathbb{R}$, when assumption 1 holds, there exists some unique critical productivity value, $\bar{q}_g$, such that a worker of group g applies to the job if and only if they have productivity $q_i \leq \bar{q}_g$.
Proof. See Appendix B for all proofs.

Corollary 1: Boundary of cut-off strategies. If a cut-off strategy exists and is defined by $\bar{q}_g$, it must be in the interval $(0, \omega)$.

With this introduced notation, I can now formally describe the best response strategies of the workers and the employer. Let $\alpha(\phi, \hat{\psi}, \hat{m}, \hat{\underline{\mathbf{s}}})$ define a best response application strategy as a mapping of the labour demand ϕ, beliefs of the employer type and signal thresholds $\hat{\underline{\mathbf{s}}} = (\hat{\underline{s}}_A, \hat{\underline{s}}_B)$ to a tuple of critical productivity values $\bar{\boldsymbol{q}} = (\bar{q}_A, \bar{q}_B)$.

Finally, the employer's best response hiring strategy can be denoted by $h(\phi, \psi, m, \hat{\bar{\boldsymbol{q}}})$ as a mapping of labour demand ϕ, the employer's type, and their belief of the workers' strategies to a tuple of signal thresholds $\underline{\mathbf{s}} = (\underline{s}_A, \underline{s}_B)$. A hiring strategy is discriminatory if $\underline{s}_A \neq \underline{s}_B$.

It is necessary to define two corollary functions from the employer's hiring strategy. First, let $\Pi(h, \alpha)$ be the profit the employer earns playing the strategy h conditional on the workers' best response application strategies α. Second, define $\beta^g(h, \alpha)$ as the share of workers from group g hired when playing the strategies h and α.

Finally, I make the following assumption of an upper bound to the hiring capacity to ensure continuous equilibrium conditions over the productivity values in the interval $(0, \omega)$. The below assumption ensures that for any wage, the share of workers who would prefer the firm job with certainty is larger than the hiring capacity.

Assumption 2: The hiring capacity of the employer is bounded by the distribution of productivity values with respect to the wages,

$$\phi \leq F(\omega) \tag{3}$$

2b Correct attribution of employer type

I first define the equilibrium when workers correctly attribute the employer to be unbiased and sophisticated. As I show in this section, this type of employer is a statistical discriminator in the Arrovian sense (Arrow, 1974): discrimination only occurs if the employer believes that workers play separating strategies.

Accordingly, under correct attribution of employer type in the job application domain, multiple equilibria may exist as is the result in Coate and Loury (1993). In this model, one of the equilibria is a non-discriminatory equilibrium in which workers play pooling strategies and the employer sets equal hiring thresholds. The other equilibria are discriminatory in which workers, perceiving potential discrimination, play separating group-dependent strategies leading the sophisticated employer to set discriminatory hiring thresholds.

In the following equilibria, beliefs are self-confirming: the strategy each side expects the other to play is precisely the strategy the other side finds optimal. Formally, workers' beliefs of hiring thresholds must generate application cut-offs that, in turn, make those thresholds profit-maximising for the employer. Additionally, the hiring strategy must satisfy two further conditions. First, if the employer sets the hiring threshold $\underline{s}_A$, it must be profit maximising to set the hiring threshold $\underline{s}_B$. In other words, the employer cannot earn more in expected profits from setting a hiring threshold $\underline{s}'_B \neq \underline{s}_B$ if it is already implementing $\underline{s}_A$. I refer to this as the *optimality* condition. Second, since the labour demand is exogenously given, for a set of strategies to constitute an equilibrium, the share of workers hired must satisfy this labour demand. I refer to this condition as one of *satiation*.

Definition 1: Given that the employer is known to be unbiased and sophisticated, a vector of beliefs of application strategies $\bar{\boldsymbol{q}} = (\bar{q}_A, \bar{q}_B)$ and signal thresholds $\underline{\boldsymbol{s}} = (\underline{s}_A, \underline{s}_B)$ constitute a Perfect Bayesian Equilibrium when beliefs of strategies are consistent such that

$$\bar{\boldsymbol{q}} = \alpha(\phi, \psi, m, h(\phi, \psi, m, \bar{\boldsymbol{q}})). \tag{4}$$

and the employer's strategy meets the following conditions:

1. Optimality. Given a signal threshold applied to one group, the employer maximises their profits by setting a signal threshold for the other group. Thus without loss of generality, for a given signal threshold $\underline{s}_A$,

$$\Pi(h(\phi, \psi, m, \bar{\boldsymbol{q}})) \geq \Pi(h'(\phi, \psi, m, \bar{\boldsymbol{q}})) \quad \forall h'(\phi, \psi, m, \bar{\boldsymbol{q}}) \neq h(\phi, \psi, m, \bar{\boldsymbol{q}}). \tag{5}$$

2. Satiation. The labour demand of the employer is satisfied,

$$\sum_{g \in \{A,B\}} n_g \beta^g(h(\phi, \psi, m, \bar{\boldsymbol{q}}), \bar{\boldsymbol{q}}) = \phi. \tag{6}$$

If Assumption 1 holds, I can apply Lemma 1 and assume that workers will play cut-off application strategies denoted by $\bar{\boldsymbol{q}} = (\bar{q}_A, \bar{q}_B)$. In equilibrium, the employer will have a correct belief of this tuple. Thus, their conditional expectation of the productivity of an applicant from group g portraying a signal s is the following,

$$\mathbb{E}[q|s, g; q \leq \bar{q}_g] = \frac{t(\bar{q}_g) \cdot u(\bar{q}_g) + \eta s}{t(\bar{q}_g) + \eta} \tag{7}$$

where $t(\bar{q}_g)$ is the concentration and and $u(\bar{q}_g)$ is the mean of the distribution of ability among candidates given that their ability is no greater than $\bar{q}_g$.

In Lemma 2, I demonstrate that if the unbiased sophisticated employer believes workers play pooling application strategies, their unique hiring strategy is to set non-discriminatory signal thresholds. In other words, an unbiased and sophisticated employer can only rationally play discriminatory hiring thresholds if they believe that workers of different groups play different strategies.

Lemma 2: Fair and sophisticated employer is an Arrovian statistical discriminator. Given Assumption 1 and that the employer of type $m = 0$ and $\psi = 1$ believes that workers play the pooling application strategies, $\hat{\bar{q}}_A = \hat{\bar{q}}_B$, the optimal hiring strategy is $\underline{s}_A = \underline{s}_B$.

Lemma 2 therefore rules out discrimination in equilibrium whenever workers pool, which means the only role for discrimination is through separating strategies.

From Lemma 1, each hiring strategy maps to a unique optimal application strategy. This means that any non-discriminatory hiring strategy can be consistent with a self-confirming equilibrium and satisfy optimality. To show the existence of a non-discriminatory equilibrium, it must be that the satiation condition can simultaneously hold for some pooling application strategy.

Proposition 1: Existence of non-discriminatory equilibrium. Suppose that Assumptions 1 and 2 hold and the employer is known to be unbiased ($m = 0$) and sophisticated ($\psi = 1$). There exists a unique non-discriminatory equilibrium.

Proposition 1 guarantees that a non-discriminatory equilibrium exists when the employer is correctly perceived to be unbiased and sophisticated. Proposition 2 establishes that, under some mild parametric restrictions, discriminatory equilibria may also exist.

Proposition 2: Existence of discriminatory equilibria. Suppose Assumptions 1 and 2 hold and the employer is known to be unbiased ($m = 0$) and sophisticated ($\psi = 1$). There exist non-empty sets $\Omega \subset \mathbb{R}_{++}$ and $\Phi \subset (0, F(\omega)]$ such that whenever $\omega \in \Omega$, and $\phi \in \Phi$, there exist two discriminatory Perfect Bayesian equilibria with application cut-offs $\bar{q}_A \neq \bar{q}_B$.

Figure 1 illustrates the existence of the multiple equilibria under correct attribution of employer type at the intersections of the two equilibrium conditions. The set Ω defines wage values such that the optimality condition admits solutions off the 45-degree line. This defines the optimality condition as being satisfied by tuples of application strategies on the 45-degree line and the closed oval locus. The set described by Φ ensures that the satiation locus intersects the 45-degree line inside the closed oval locus defined by the optimality condition. The non-discriminatory equilibrium (NDE) lies on the 45-degree line where the application strategies for each group are equal. The two discriminatory equilibria (DE1 and DE2) lie off this line. In these equilibria, one group's applicant pool has a higher average productivity than the other, meaning that conditional on a signal value, the expected productivity of an applicant from one group exceeds that of the other, providing a rational basis for the employer to set differential hiring thresholds.

Corollary 2: If Assumption 1 holds, the equilibrium earnings of a group g is increasing in their critical productivity value $\bar{q}_g$.

By Corollary 2, the worker group who is discriminated against in a discriminated equilibrium is worse off than the group which is favoured. Moreover, the discriminated group would prefer the non-discriminatory equilibrium from a Utilitarian welfare perspective than a discriminatory equilibrium against their favour.

2c Misattribution of employer type

In this section, I consider the case where workers expect the employer to harbour prejudicial preferences, $\hat{m} \neq 0$. By symmetry of groups A and B, it is without loss of generality to consider that workers expect the bias to be in favour of group B, $\hat{m} < 0$.[11] Therefore, the ensuing equilibrium

[11]This is just a labelling convention: the case where $\hat{m} > 0$ is symmetric and yields analogous results. In the extreme case where $\hat{m} \to -\infty$, the employer is perceived to have infinite taste for B workers. In this

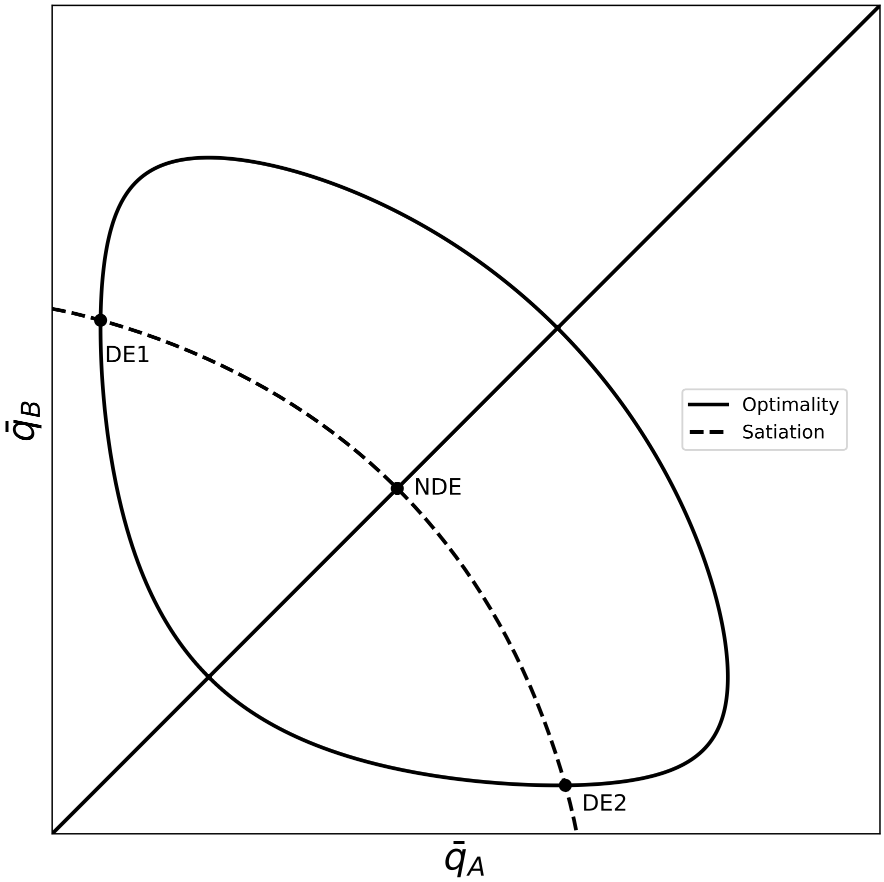


Figure 1: Perfect Bayesian equilibria under correct attribution

Note: Example of multiple Perfect Bayesian equilibria existing when workers correctly believe the employer to be fair and sophisticated. Along the axes are the application strategies of each group: the maximum productivity value at which an individual from group $g \in \{A, B\}$ would apply to a job. The loci labelled *Optimality* plots the combination of application strategies that are rationalised through consistent beliefs which lead to hiring strategies which jointly satisfy unconstrained profit maximisation (equilibrium condition 1). The locus labelled as *Satiation* plots the combination of self-confirming application strategies that lead to hiring strategies which jointly satisfy labour demand (equilibrium condition 2). The intersections of these loci constitute the equilibria. The intersection marked by NDE is the non-discriminatory equilibrium. The point DE1 (DE2) is the discriminatory equilibrium which favours group B (A).

definition under misattribution must be adjusted to account for inconsistent beliefs of the employer's type while still ensuring consistent beliefs of strategies. Given the misspecified beliefs of the workers, this now constitutes a Bayesian Nash Equilibrium. In this equilibrium, workers form beliefs of the thresholds based on their (mis)perception of the employer type, but the thresholds actually implemented are those of the true type.

Definition 2: A vector of beliefs of application strategies $\bar{\boldsymbol{q}} = (\bar{q}_A, \bar{q}_B)$, signal thresholds $\underline{\boldsymbol{s}} = (\underline{s}_A, \underline{s}_B)$ and beliefs of employer type $(\hat{m}, \hat{\psi})$ constitute a Bayesian Nash Equilibrium when, given

case, A workers will perceive that their probability of being hired is 0 and we arrive at a corner solution where $\bar{q}_A = 0$.

beliefs of employer type, strategies are consistent such that

$$\bar{\boldsymbol{q}} = \alpha(\phi; \hat{\psi}, \hat{m}, h(\phi, 0, 1, \bar{\boldsymbol{q}})), \tag{8}$$

and that, given this $\bar{\boldsymbol{q}}$, the employer's strategy meets conditions 1 and 2 in Definition 1.

Proposition 3 demonstrates that the non-discriminatory equilibrium is not perceived as attainable when the employer is perceived as biased $\hat{m} < 0$.

Proposition 3: Perceived non-existence of non-discriminatory equilibria under misattribution. Suppose Assumptions 1 and 2 hold. The non-discriminatory equilibrium is not a Bayesian Nash Equilibrium under Definition 2 when $\hat{m} < 0$ for any $\phi \in (0, F(\omega)]$.

Intuitively, Proposition 3 demonstrates that in order for the non-discriminatory equilibrium to prevail, workers must correctly attribute the employer to be unbiased: that they are not a taste-based discriminator. Proposition 1 establishes that a non-discriminatory equilibrium exists under correct attribution; Proposition 3 shows that misattribution of employer type removes this possibility from the perceived set of equilibria.

For simplification, I herein assume that the workers also perceive the employer to be naïve ($\hat{\psi} = 0$). Intuitively, the assumption of supposed naïvety means that workers expect the employer to believe in equal application strategies $\bar{q}_A = \bar{q}_B$, and therefore will not infer different underlying productivities from equal signals of applicants holding different identities.[12] From this misattribution, workers will expect the optimality and satiation rules to which the employer adheres to be $\hat{\underline{s}}_A = \hat{\underline{s}}_B - \hat{m}$ and $\phi = n_A \beta^A(\hat{\underline{s}}_A, \bar{q}_A) + (1 - n_A)\beta^B(\hat{\underline{s}}_A + \hat{m}, \bar{q}_B)$ respectively. Because workers believe the employer ignores application strategies, they perceive the employer's profit-maximisation rule as a direct mapping from the perceived bias $\hat{m}$ to a unique pair of hiring thresholds. This feature eliminates the multiplicity from the workers' perceived strategy space that otherwise generates the non-discriminatory equilibrium. Importantly, misattribution entails the workers' misperception of the employer's optimality condition.

In the rest of this section, I demonstrate that misattribution may only sustain a discriminatory equilibrium, and given a belief of the employers' preferences, this equilibrium is observationally equivalent to the one derived in Proposition 2. This result arises not from the employer's problem, which is unchanged, but from the workers' misperception of the strategy space available to them.

The following outlines that individuals may encounter discrimination of a purely statistical nature but may incorrectly believe that this is caused by a taste-based discriminator and the strategies observed sustain both the belief that discrimination is occurring and the belief about its source.

Proposition 4: Misattribution equilibrium. Suppose that Assumption 1 holds, $m = 0$, $\psi = 1$ and beliefs are $\hat{\psi} = 0$ and $\hat{m} < 0$. Moreover, suppose that the parametric conditions of Proposition 2 are satisfied, i.e. $\omega \in \Omega$, $\phi \in \Phi$. Then there exists some $\hat{m} < 0$ such that a discriminatory Bayesian Nash Equilibrium exists that is observationally equivalent to the discriminatory Perfect Bayesian Equilibrium established in Proposition 2: the equilibrium application cut-offs $(\bar{q}_A, \bar{q}_B)$ and corresponding hiring thresholds $(\underline{s}_A, \underline{s}_B)$ coincide.

Crucially, from Proposition 4, if workers perceive the employer to be biased with some preferences $\hat{m} < 0$, then their optimal strategy can lead to the same discriminatory equilibrium as that

[12]This assumption results in a unique equilibrium under misattribution which is useful for computation however it does not affect the implications of the main result of the model. Relaxing the restriction expands the set of equilibrium, however, by Proposition 3, this cannot include a non-discriminatory equilibrium.

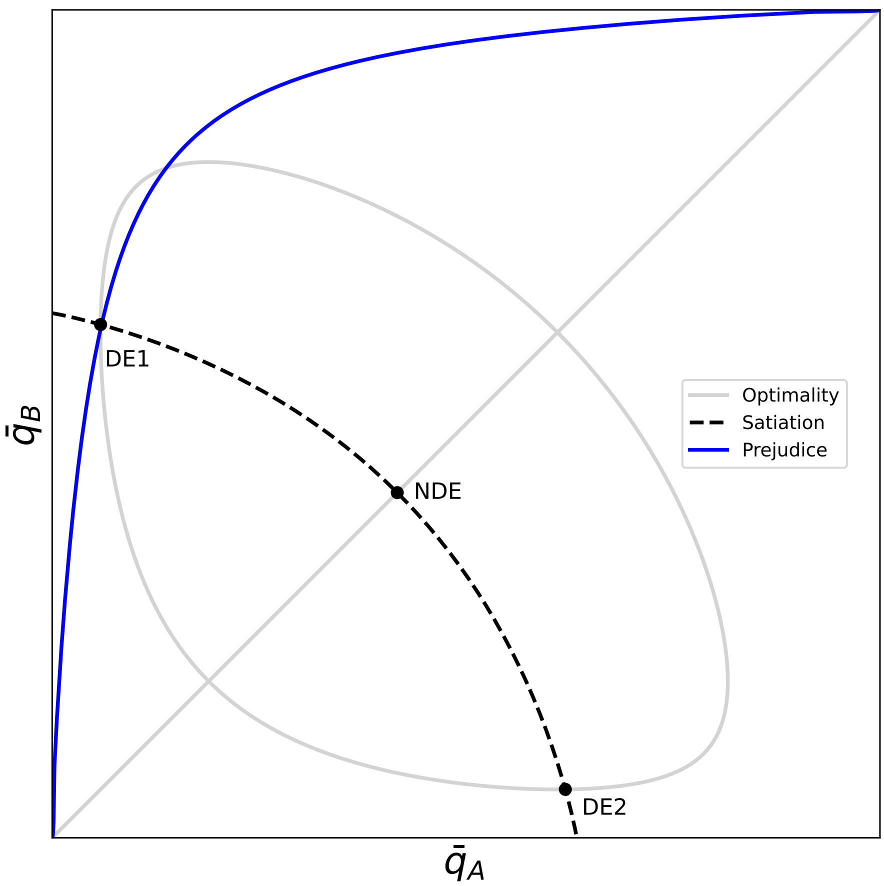


Figure 2: Bayesian Nash equilibrium under misattribution

Note: Example of Bayesian Nash equilibria under misattribution: incorrect beliefs of employer type. Along the axes are the application strategies of each group: the maximum productivity value at which an individual from group $g \in \{A, B\}$ would apply to a job. The loci labelled *Optimality* plots the combination of application strategies that lead to hiring strategies which jointly satisfy unconstrained profit maximisation (PBE condition 1). The locus labelled as *Satiation* plots the combination of application strategies that lead to hiring strategies which jointly satisfy labour demand (PBE condition 2). The locus *Prejudice* maps the application strategies that are incentivised by self-confirming beliefs of a naïve and biased employer of preference $\hat{m} < 0$. The intersection of the *Satiation* and *Prejudice* loci (DE1) is the only attainable equilibrium in this setting since workers perceive the employer's profit maximising condition to be the *Prejudice* locus rather than the *Optimality* locus. Although under different beliefs, the other PBE equilibria (NDE and DE2) are attainable as in Figure 1.

under correct attribution of employer type. In this case, misattribution leads to an equilibrium that does not challenge the beliefs of the employer type. Since the non-discriminatory equilibrium does not exist and that equilibrium coordination is determined by beliefs of others' strategies, there is no incentive to coordinate, either organically or through policy directives, on a different equilibrium. Additionally, Proposition 4 suggests a potential mechanism through which misattribution may persist: that workers misperceive the sophistication of the employer leading them to misperceive the set of equilibria. In Appendix C, I derive an alternative mechanism: that agents are aware of the sophistication levels of the employer but misperceive the strategies of other agents.

In Figure 2, I include the locus of beliefs of application strategies that are self-confirmed by biased and naïve hiring thresholds for some value $\hat{m} < 0$. By naïvety, the employer is believed to ignore application strategies, so the locus is injective and each $\hat{m}$ maps to a unique threshold ratio. As $\hat{m} < 0$, this locus lies off the 45-degree line. The satiation locus remains unchanged from Figure 1. In the case of misattribution, agents incorrectly perceive the employer's profit maximising condition as the locus labelled *Prejudice* instead of as the *Optimality* loci. The intersection of *Prejudice* and *Satiation* is then the only perceived equilibrium which is the same as the rational discriminatory equilibrium (DE1). This demonstrates the misattribution equilibrium: where workers' best response to a belief of taste-based discrimination triggers a hiring strategy from a fair but sophisticated employer that reinforces workers' beliefs.

2d Applications

There are two key interpretations of the preceding model of misattribution. First, the more workers overweight the probability that the employer is a taste-based discriminator, the less attainable a non-discriminatory equilibrium appears, and the more likely they are to coordinate on a discriminatory one. Second, because the discriminatory equilibria under correct and incorrect attribution are observationally equivalent, workers cannot learn the employer's true type by observing discriminatory outcomes alone. Determining the employer's type in this case requires costly coordinated experimentation, a cost that misattribution inflates by making such a trial appear futile. Taken together, this section derives a testable predictions: if the probability that the employer is a taste-based discriminator is overweighted, individuals will be more likely to play application strategies that sustain the underlying statistical discrimination. Discrimination is not only sustained by self-confirming beliefs, but its perceived source determines whether workers ever generate the evidence needed to break the trap. In the proceeding sections, I outline a laboratory experiment in which I test these predictions.

This channel is plausible outside the laboratory. Using data from questions in Gallup World Poll (2025), it appears that while in the not-too-distant past, bias or prejudicial preferences were common but have since all but disappeared. In 1958, 94% of surveyed Americans disapproved of interracial marriages, whereas only 4% did so in 2021. Similarly, 64% of American respondents stated that they would not vote for a well-qualified female presidential nominee from their party in 1937; in 2024, that figure stood at 5%.[13] However, in a representative survey carried out in 2025, I find that participants overestimate the share of others who held such prejudicial views (see Appendix A for details). Concurrently, individuals' beliefs of the presence of discrimination have remained stable or increased over the recent decades.[14] This divergence is consistent with the feedback mechanism the model implies: if individuals misperceive the probability of encountering a taste-based discriminator, they take actions that force rationally sophisticated but informationally constrained employers to discriminate, sustaining the belief that generated them.

[13] These questions have been used as proxies for distaste against minority groups (see e.g., Charles and Guryan, 2008)

[14] When asked *In general, do you think that Black people have as good a chance as White people in your community to get any kind of job for which they are qualified, or don't you think they have as good a chance?*, 27% of respondents to the Gallup World Poll stated 'No, do not' in 2008 whereas 42% stated the same in 2025. Regarding men and women, 42% of respondents stated that they believe women do not have equal job opportunities as men in 2008 whereas 48% stated this belief in 2025 (Gallup, 2025).

3 Laboratory experiment design

3a Set-up

The experiment is designed to mirror the decision making process outlined in the previous section to test for the dynamic stability implied by the static model. Although the model itself is static, this stability hinges on how beliefs about the employer's type update over time which requires repeated actions and signal structures to observe.

To allow for the possibility of misattribution, I present participants with an ambiguous hiring procedure in which neither the employer's evaluation method nor the application decisions of others are observable. To achieve the former, I describe two types of employers that may make employment decisions. One type is chosen at the beginning of the session by a random draw and participants attribute outcomes to a type by stating their subjective probability that the employer is one of the binary types. To establish strategic uncertainty, multiple participants simultaneously make application decisions for a set number of jobs in a competitive labour market.

Structure. I recruited sessions of 32 participants who were initially randomised into an experimental *society*. Each *society* contained 16 participants who only interacted with others in their *society* for the entirety of the experiment. Once formed, I randomly allocated members of each society into two equally large identities: green and purple. I chose to use minimal identities to shield participants from influences from real markets as this has been shown to influence beliefs of discrimination and behaviours (Ruebeck, 2024; Gagnon et al., 2025). Green and purple are used to separate colours from any real affiliations such as politics or gender as in Dianat et al. (2022).

In each round played, participants were assigned a productivity value from a known normal distribution of mean 50 and standard deviation 81.85, and had to make 3 incentivised decisions.[15] The first decision involved a labour allocation choice: choosing to apply for one of 3 available jobs paying 90 tokens, or to earn an income equal to one's drawn productivity for the round. An unsuccessful application led to a round income of 0 tokens. The employer, who made the hiring decisions, could observe the identity of the applicant but was unable to observe their true productivity. Instead, they observed a noisy but unbiased signal of such computed as the sum of the participant's productivity value and an idiosyncratic noise value drawn from a zero-mean normal distribution with standard deviation 81.85.[16]

The employer was described to be one of two types, A or B.[17] Before the session began, one type of employer was randomly drawn and implemented for all members of a society in all rounds of the experiment. Participants were informed that type A was drawn with probability $p > 0$ and type B with probability $1 - p$. The type A employer served as my biased naïve decision maker: the taste-based discriminator. This employer was described as having a fixed bias against one group: it first subtracted 90 from each signal coming from a green applicant, added 10 to each from a purple applicant, and hired the 3 applicants with the highest remaining signal values. Type B was my Arrovian discriminator which was described to predict the likely distribution of productivity

[15]These parameters were calibrated to fit the existence of multiple equilibria and a misattribution equilibrium.

[16]To explain the random draw procedures, I presented participants with two urns: one for productivity values and one for signal noise. Each urn contained 1000 balls, each with an integer written on it. Half of the balls possess a number below the mean and half at or above it. Negative numbers exist. Participants were presented with a sliding tool with which they could explore how many of the balls in the urn have a number higher and lower than different values. Participants were informed that if they applied for a job, their signal would equal the sum of the integers drawn from the two urns.

[17]To ensure deterministic employment decisions and control for expectations of such, I described these types as different robots to which the true employer had delegated the hiring responsibility.

among applicants from each identity using all available information (including past rounds), and estimate the true productivity of each applicant using this information.[18]

After making their application decision, participants were led to decision 2, in which they were asked to guess the maximum productivity value of other participants from each coloured identity in their society who chose to apply in the current round.[19]

Throughout all rounds, participants could observe a running history of hiring outcomes by identity group in graphical form. Once all participants in the society had completed decision 2, the hiring decisions were made by the employer and the hiring history graphic was updated.

Next, participants proceeded to decision 3 in which they were asked to state their belief of the probability that each type of employer which was making hiring decisions for their society. I refer to this belief as the *subjective probability of employer type*.[20]

Parametrisation of the experiment. I pre-registered the value of p, the probability that the employer was type A, to be 0.01. Participants were unaware of this value. In the week prior to the experiment, I drew 24 random integers between 1 and 100. For every time the number 1 was drawn, I would have run a session with the type A employer. All of the drawn integers were greater than 1 and so all sessions proceeded with employer type B.

A type B employer requires an initial prior of the maximum productivity of green and purple applicants, from which it bases the first round of hiring decisions and the updated belief of application strategies in proceeding rounds. I preregistered the initial prior to be at mid-point of the non-discriminatory and discriminatory (in favour of purple participants) equilibria. Effectively, employer type B *believed* that green (purple) participants would only apply if and only if their productivity value was less than 19.75 (51.5) in round 1.

3b Treatments

The above describes the baseline procedure which will be used to test for the presence of misattribution and its direct effect on leading to application strategies that perpetuate discrimination. I employed two additional treatments randomly assigned on the society level such that each treatment and the baseline had an equal number of societies and subjects.

The *revealed* treatment proceeded identically to the baseline up to and including round 4 and so I pool these rounds together with the baseline to create a baseline set of data. After round 4, participants in the *revealed* treatment were informed that the true employer was type B. The intention of this treatment is to generate exogenous variation in beliefs of employer type; holding the history of hiring outcomes fixed; in order to identify the causal effect of attribution on survey responses and application behaviour.

The *strategic* treatment was implemented to reduce the degree of strategic uncertainty in the game in order to test whether participants' misperception of others' strategies leads to misattribution. I relegate discussion of this treatment and the results to Appendix C.

[18]See Appendix D for a description of the practical procedure of the updating performed by type B.

[19]A binary scoring rule was employed for each guess wherein participants receive 50 tokens if they were within 10 of the true value, 25 tokens if they were within 20 of the true value and nothing otherwise. If no other participants of that colour applied in the round, participants were informed that their earning would be 0. Importantly, because I explicitly stated that the belief refers to others in their society and not themselves, participants could not individually force this decision to payout.

[20]A quadratic scoring rule was employed for this guess (Brier, 1950).

3c Procedural details

The experiment was conducted in the CeDEx laboratory at the University of Nottingham in November 2025 over three consecutive days and 12 sessions.[21] Participants were recruited using ORSEE (Greiner, 2015) and the experiment was coded using oTree (Chen et al., 2016). Each session consisted of 32 participants split into two equally sized societies. Thus, there were in total 24 societies amounting to 384 participants.

Before commencing the rounds, the instructions presented in Appendix G were read out loud to participants. They had to answer 2 sets of questions to assert their comprehension before proceeding. Additionally, a physical summary sheet was provided to each participant.

The monetary value of each token was set to £0.03. For each decision (1, 2 and 3), an independently drawn random round determined payment. There was also a survey conducted after the conclusion of rounds, during which subjects could earn additional bonuses. The entire experiment lasted approximately 70 minutes and participants were paid an average of £17.08 which included a show-up fee of £8.00.

4 Results

The analyses follow the pre-analysis plan with a few exceptions outlined in Appendix F. The data from the experiment is structured in the following way. For each individual i of society s in session j, I have multiple observations over periods t. I will report results using the entire set of data, excluding the first round where appropriate. Since treatment is allocated at the society level, I cluster standard errors by society.

4a The presence of misattribution in baseline data

Preliminaries. The outcomes of the experiment are determined both by the decisions of participants and also by the series of random draws. While the distribution of productivity values and noise draws are independent across sessions, initial allocations and decisions have lasting consequences. This means that not all societies in every round may experience discrimination to be attributed.

To test whether discrimination is correctly attributed, I begin by defining three clusters of hiring histories which can broadly define the state of discrimination in societies. I refer to the instance of a colour group being favoured as when the current cumulative share of hires going to that group is at least two thirds of all hiring decisions, and the instance of no discrimination being when neither group is favoured. Using these definitions, 75% of participants in the baseline data experience the purple group being favoured at some point. However, only in 50% of participant-period observations does this occur. As presented in Figure 3, societies converge to either favour purple applicants or favour neither purple nor green applicants.

A concern regarding the experiment is the attention of subjects. I define a participant as inattentive if they make a globally dominated decision: an application decision that is rationally dominated by any subjective probability of the employer type. By the theoretical model, regardless of one's belief of the employer's type, it is irrational to apply for a job if one's drawn productivity value is 90 or greater as the expected payment from application is less than 90 and the certain payment from not applying is at least 90. It is also irrational to not apply for a job if one's drawn

[21] A pilot of two sessions was conducted the day before the main sessions began. At this point, it was determined that, due to timing and financial constraints, only 8 rounds could be conducted instead of the planned 12. Changes were implemented overnight in order to satisfy the laboratory's needs.

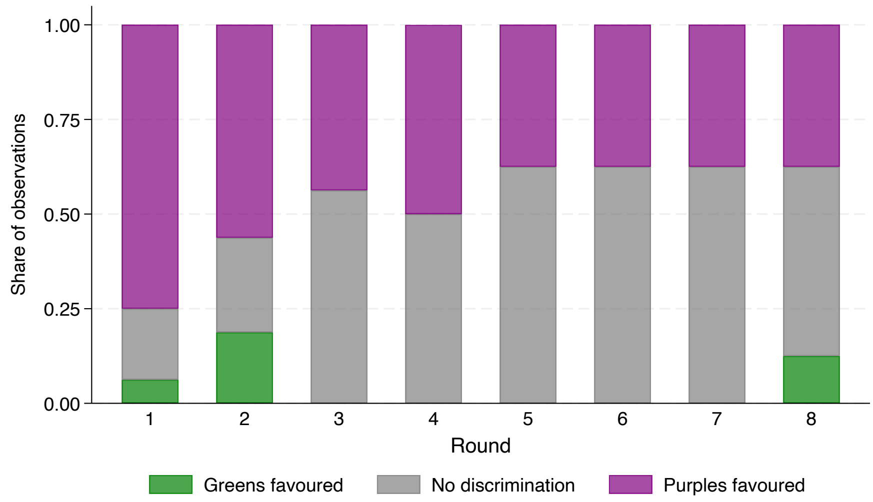


Figure 3: Share of favouritism across rounds in the baseline

Note: This graph shows the distribution of the cumulative average of hiring decisions in each round in the baseline data. "Greens favoured" refers to the occasions where more than 2/3 of the past hiring decisions have been to green workers. "No discrim" refers to occasions where this cumulative average shows that the share of past hiring decisions in favour of purple workers has been between 1/3 and 2/3. "Purples favoured" refers to the occasions where 2/3 or more of the past hiring decisions have been to purple workers.

productivity value is less than 0 as this would guarantee a negative payment versus a positive expected payment from applying.

In total, 39 (15.23%) individuals violate these rationality rules at least once, however, after round 1, only 9 (3.52%) do so. Moreover, of the observations in which the drawn productivity value is outside of the range [0,90), merely 6.23% behave irrationally. Importantly, the prevalence of irrational decisions does not differ by the favouritism of the hiring decisions as is shown in Figure E.2 (joint F-test $p = 0.813$). I report robustness tests which exclude participants who violate the rationality rules.

Beliefs of employer type. To test whether participants misattribute the source of discrimination, I report the average subjective probability that the employer is type A – a taste-based discriminator – when subjects face the expected discrimination: a hiring history in which purples are favoured. One of the key interpretations of Proposition 4 is that, at any given level of discrimination in the expected direction (against green workers), the identification problem persists: one cannot attribute taste-based discrimination from statistical discrimination. Therefore, I compare the subjective probability to a Bayesian benchmark of 50%. This benchmark is consistent with uniform priors over the drawn value of p. Figure 4 presents the cumulative distributions of beliefs at different clusters of hiring histories which are significantly distinguishable from each other (all pair-wise Kolmogorov-Smirnoff tests $p < 0.001$).

On average, subjects experiencing hiring histories that favour purple applicants believe that the probability of the employer is of type A is 57.0%. This is significantly above the Bayesian benchmark of 50% as shown in Figure 5 (Wilcoxon signed-rank test $z = 7.78$ $p < 0.001$). This result is robust to the exclusion of observations from round 1 (average belief 54.8%; Wilcoxon signed-rank test vs. 50% $z = 5.16$ $p < 0.001$).[22]

[22]I separate this analysis by round in Figure E.3. In all rounds where purples are favoured, the average belief is above the Bayesian benchmark of 50%. I am not powered to test the statistical significance of the

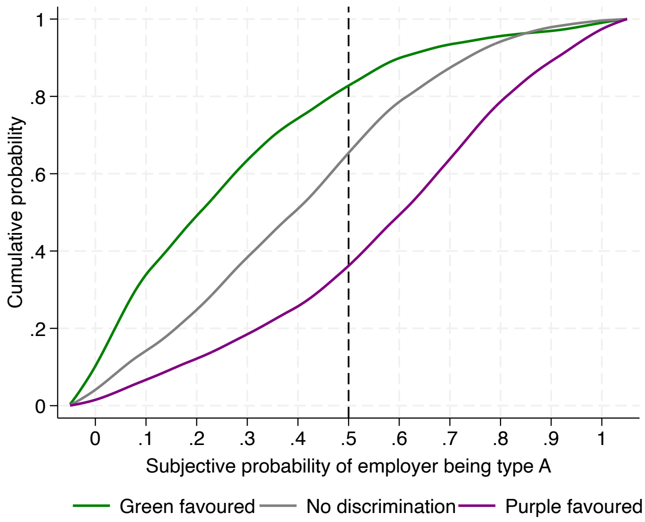


Figure 4: Cumulative distribution of subjective probabilities that the employer is type A, by favouritism levels

Note: This graph shows the cumulative distributions of the subjective probability that the employer is type A – the taste-based discriminator – at different clusters of hiring histories. "Greens favoured" refers to the occasions where more than 2/3 of the past hiring decisions have been to green workers. "No discrim" refers to occasions where this cumulative average shows that the share of past hiring decisions in favour of purple workers has been between 1/3 and 2/3. "Purples favoured" refers to the occasions where 2/3 or more of the past hiring decisions have been to purple workers. The horizontal line represents the Bayesian benchmark belief when greens are not favoured.

Econometric analysis. In the above, the definition of favouritism and thus discrimination is arbitrary. To test if the cumulative share of hires going to purple workers affects the subjective probability of employer type, I estimate a linear probability model of the subjective probability that the employer is type A on the cumulative average share of hires going to purple participants. The estimates are presented in Table 1. In these regressions, the more in favour a hiring history is towards purple workers, the higher the subjective probability that the employer is type A. When including session, individual and round fixed effects, a 10 percentage point increase in share of historical hires that are in favour of purple increases the average subjective probability of the employer being type A by 8.54 percentage points. This coefficient estimate is significant at the 1% level of significance.[23] Thus, the result can be summarised below.

Result 1: Participants who experience discrimination misattribute the source of discrimination to a taste-based discriminator. The magnitude of misattribution increases with the intensity of the discrimination.

4b Misattribution directly perpetuates discrimination

In this subsection, I test the main hypothesis derived from the theoretical model: that participants who misattribute the source of discrimination are more likely to make application behaviour con-

difference for each individual round.

[23]This result is robust to the exclusion of round 1 estimates (see Table E.1), irrational decision makers (see Table E.2) and to the analysis of round pairings in isolation (see Table E.3).

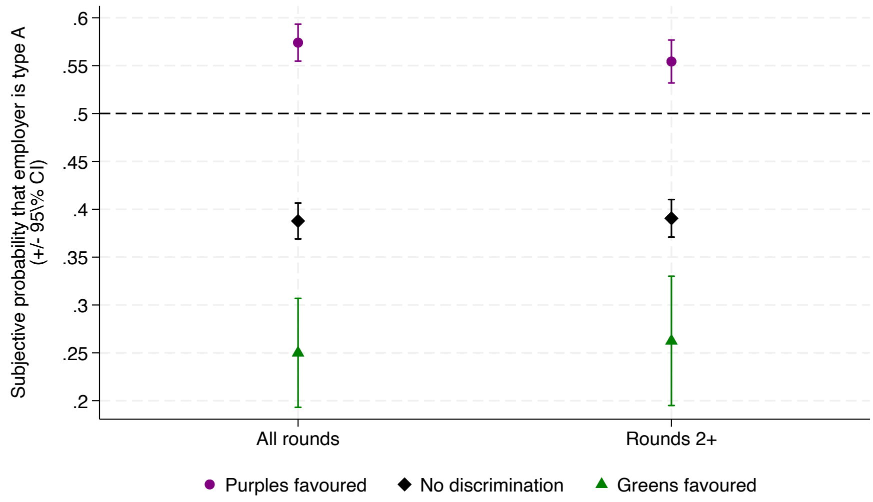


Figure 5: Average subjective probability that the employer is type A by discrimination incidence

Note: This graph shows the average subjective probability that the employer is type A (the taste-based discriminator) is in use by individuals experiencing different levels of discrimination (proxied by the cumulative average share of hires for purple). Purples (Greens) favoured refers to the case where the cumulative average share of purple (green) hires is at least 2. No discrimination refers to cases in between these bounds. Average beliefs are shown for all rounds and those excluding round 1. Error bars show 95% confidence intervals.

Table 1: Misattribution: Effect of cumulative average share of purple hires on believed probability of the employer being type A

Dependent Variable:	Subjective probability that employer is type A			
	(1)	(2)	(3)	(4)
Share of hires purple	0.607***	0.692***	0.850***	0.854***
	(0.080)	(0.069)	(0.100)	(0.120)
Session FEs		✓	✓	✓
Individual FEs			✓	✓
Round FEs				✓
Baseline mean	0.474	0.474	0.474	0.474
Observations	1536	1536	1536	1536
R^2	0.200	0.243	0.561	0.565

Notes: Linear fixed effects model estimation of the subjective probability that the employer is type A, the taste-based discriminator, on the cumulative share of previous hires that go to purple candidates. Sample includes the baseline treatment and first four rounds of *revealed* treatment. Baseline mean is the average subjective probability the employer is type A over the whole sample included where the average share of hires going to purple candidates was 60.5%. Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

gruent with a discriminatory equilibrium. To do so, I restrict my sample to observations wherein the participant draws a productivity value in the interval $[0, 90)$. This is to ensure that participants included in the analysis may rationally choose either application action depending on their belief of the hiring threshold.[24]

In this subsample, some individuals are only observed in a small number of rounds leaving insufficient within-person variation for the individual fixed-effects estimator to identify their contribution. Instead, I employ a pre-specified bank of control variables that are collected in the experiment including gender, age, education level, student status, economic and social political leaning and risk preference. As my dependent variable of interest is a binary action, I estimate a logistic fixed effects model of the application decision on the previous round's stated subjective probability of the employer being type A. Using this model, I can test whether misattribution distorts application decisions conditional on some level of discrimination observed and current productivity.[25] The effect on application leads to sustained discrimination if the impact of misattribution is heterogeneous across productivity values. Specifically, one would expect that the largest difference in application behaviour responses to misattribution to be for individuals with productivity values in the middle of the included set. To test for this heterogeneous effect, I therefore include a quadratic term for productivity value draws in the regression model.

Table 2 shows the estimated coefficients from this model. The results show that the higher the stated subjective probability that the employer is type A from the previous round, the less likely green participants are to apply and the more likely purple participants are to apply in the current round. Computing the average marginal effects using specification 4, a 10 percentage point increase in the subjective probability that the employer is type A leads to a 3.5 percentage point decrease in the probability that a green worker with productivity in the range $[0, 90)$ will apply to the job and a 2.4 percentage point increase in the probability that a purple worker with productivity in this range will apply.[26] This demonstrates that the higher the subjective probability that the employer is a taste-based discriminator, the difference between application propensities of discriminated and favoured workers increases creating grounds for statistical discrimination.

This behavioural effect is likely heterogeneous across participants with different productivity values as the relative attractiveness of each application action varies according to this draw. Subjects with productivity values close to the boundaries of the set $[0, 90)$ are unlikely to be responsive to their belief of employer type in their application decision. Those with low productivity values are likely to always apply to the job as their certain payment from not applying is small compared to the expected value from applying. Moreover, those with high productivity values are unlikely to apply to the job regardless of their belief since their certain payment from not applying is relatively attractive. As outlined in Section 2, the mechanism through which discrimination persists is that which sustains a difference in the distributions of productivity among the green and purple applicants. Accordingly, that misattribution directly leads to sustained discrimination relies on the effect identified above to be heterogeneous across productivity values.

Figure 6 shows that the marginal effect of the subjective probability of a type A employer on

[24]The previous result is not affected by this restriction. On average, participants in this sample believe the probability that the employer is type A is 55.3% (vs. 50% Wilcoxon signed-rank test $z = 15.6$ $p < 0.001$). See Table E.4 for a replication of Table 1 using this restricted data.

[25]As the independent variable is endogenous, the results from this section should not be interpreted as causal claims. To study causality, a larger sample will be necessary to study the treatment effect in round 5 of the exogenous variation in beliefs on application propensity.

[26]This effect is robust to the exclusion of the first decision following a hiring history feedback (see Table E.5 columns 1-4). When excluding the final four rounds to account for learning, the estimated coefficients are in the same direction as the full specification but are no longer statistically significant due to power limitations (see Table E.5 columns 5-8).

Table 2: Misattribution perpetuates discrimination: Effect of subjective probability of the employer being type A

Dependent Variable:	Application decision			
	(1)	(2)	(3)	(4)
Subj. prob of type A (Previous round)	-1.420* (0.736)	-1.802** (0.739)	-2.005** (0.854)	-2.068** (0.849)
Purple	-1.384** (0.546)	-1.483** (0.592)	-1.484** (0.586)	-1.504*** (0.554)
Purple × Subj. prob of type A (Previous round)	2.660** (1.172)	2.884** (1.284)	3.294** (1.402)	3.402** (1.357)
Effect of belief on purple	1.240*	1.082	1.289*	1.334**
Session FEs		✓	✓	✓
Individual Controls			✓	✓
Round FEs				✓
Baseline mean	0.539	0.539	0.539	0.539
Observations	381	381	381	381
R^2	0.147	0.181	0.229	0.241

Notes: Logistic fixed effects models of the propensity of application on the subjective probability that the employer is type A. The independent variable is interacted with an indicator of whether the participant is purple colour to show the heterogeneous effects on application for each colour. All specifications include a quadratic control for productivity level and a linear control for the cumulative share of hires going to purple workers from the previous round (the hiring history observable when the participant makes their application decision). Sample is restricted to participants who draw a round productivity value in the interval $[0, 90)$. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

the propensity to apply is strongest for workers with productivity values around 45. In fact, there is no significant marginal effect for those with productivity values of 0. Using the estimated average marginal effects, for a green (purple) worker of productivity value 45, a 10 percentage point increase in the believed probability of a taste-based discriminator being used leads to a decrease (increase) in their probability of applying by approximately 4.1 (2.8) percentage points.[27] An interpretation of this result is that the more probable participants perceive the employer to be type A, the lower the expected productivity of a green applicant and the higher the expected productivity of a purple applicant.

Result 2: Misattribution perpetuates statistical discrimination by reducing the probability of ap-

[27] I plot the adjusted predicted probabilities of application at 2 different levels of attribution for each colour in Figure E.4: a certain attribution of employer type A and a certain attribution of employer type B. For green workers (panel (i)), the less probable the employer is type A the more likely is an application. The reverse is true for purple workers (panel (ii)). For both colour groups, the effect is strongest for participants who draw productivity values around 50.

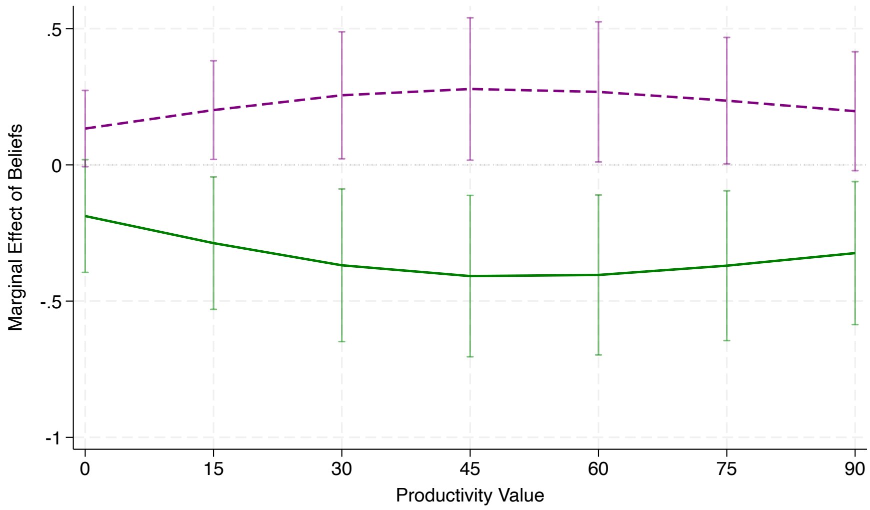


Figure 6: Marginal effect of believed probability of the employer being type A on absolute probability of application at different productivity values

Note: This graph shows the marginal effect of believed probability of the employer being type A on the propensity of a participant to apply for different values of productivity for each participant colour. Computations are based on specification (4) from Table 2. Confidence intervals are at the 95% level of significance.

plication for the discriminated group and increasing that for the favoured group.

4c Misattribution indirectly perpetuates discrimination

In the treatment arm, *revealed*, participants were informed that the true employer type was B after round 4. At this point in all sessions, a midline survey was implemented to test for indirect effects of misattribution on the persistence of discrimination. Specifically, this survey measured subjects' perception of the fairness of the experimental labour market and their demand and perceived efficacy of a temporary affirmative action.

Fairness views. Subjects reported to what extent, on a 5-point Likert scale, they considered the hiring procedure to be fair thus far. At the end of the experiment, I repeated the fairness question once more. Such evaluations can affect the stickiness of a discriminatory equilibrium: it has been shown that workers may *backlash* due to perceived unfairness in the labour market through withdrawing their effort or labour supply (Charness and Levine, 2000; Offerman, 2002; Abeler et al., 2010; Breza et al., 2018). Particularly, this question is motivated by the results of Kuhn et al. (forthcoming) who do not find average differences between the perceived fairness of taste-based and statistical discrimination but do find that the perceived justification of the discrimination matters. Applied to my setting, taste-based discrimination is parametrised according to an inherent individual bias within the model, the least justifiable taste-based discrimination found by Kuhn et al. (forthcoming). Meanwhile, the statistical discrimination is established by the behaviours of all participants in the experimental labour market across all cumulative periods: the most fair justification according to the authors' findings. If misperceiving the statistical discrimination as taste-based discrimination increases perceptions of market unfairness, an indirect effect of such could be that participants are less likely to apply to a job. This would be particularly problematic in my setting if the discriminated group perceives the market as more unfair than the favoured group at a given level of discrimination and beliefs.

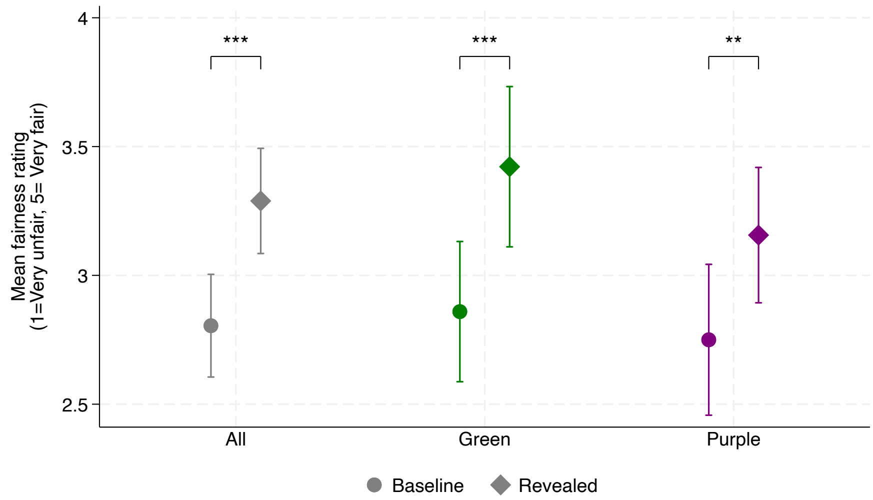


Figure 7: Fairness perecpetion of hiring process after round 4

Note: This graph shows the average response to the question *To what extent do you consider the hiring procedure has been fair thus far?* asked after round 4. Responses are coded to a 5-point Likert scale where 1 is *very unfair* and 5 is *very fair*. Individuals are asked this question after being informed of the true employer type, if in *revealed* treatment. Error bars show 95% confidence intervals. Stars show the significance of the difference computed from a two-sided t-test: *p<0.10, **p<0.05, ***p<0.01.

Comparing the baseline and *revealed* treatments reveals that subjects, on average, perceived the hiring process to be more fair when they knew the employer type was type B (Likert scale = 3.289) than when they were unsure of the employer type (Likert scale = 2.805). The difference between these is statistically significant at the 1% level (two-sided t-test $t = 3.328$ $p = 0.001$). As demonstrated in Figure 7, this difference is observable among both colour identities, which suggests that both victims and beneficiaries perceive statistical discrimination as more fair than ambiguous discrimination. This result is more striking when analysing participants' response to the same question in an final survey after all rounds are completed as shown in Figure E.5.

To analyse whether this is driven by beliefs of employer type or other treatment related differences, I run a linear regression model of participant responses to the midway fairness question on their subjective probability that the employer is type A as recorded in round 4. To generate a reliable comparison, I exclude participants in societies which, in round 4, observe less than 50% of the previous hires going to green. As the dependent variable is measured at one instance, I include session fixed effects and individual controls. I control for the hiring history to analyse how beliefs of the source of discrimination affect fairness perceptions at a given level of realised inequality.

I present the results of this estimation in Table 3. In columns 1 to 3, I include participants only from the baseline treatment arm, for whom the source of discrimination is uncertain. By the estimated coefficients, conditional on a level of discrimination, subjects perceived the hiring process as less fair the more likely they perceived the employer to be a taste-based discriminator. However, the marginal effect of the belief of employer type is only significant for green workers, the discriminated group, for whom the estimated effect size is eight times as large as that among purple workers, the favoured group. Interpreting the coefficient in specification 3, transitioning from the Bayesian benchmark perceived probability of the employer being type A of 50% to 100% reduces perceived fairness by 0.527 points on the 5-point Likert scale or by 0.458 of a standard deviation.

By hard coding the perceived probability that the employer is type A to 0% in the *revealed*

Table 3: Effect of misattribution on fairness perceptions

Dependent Variable:	Fairness view of hiring process at midline (5-point Likert scale, 1 = Very Unfair and 5 = Very Fair)					
	Baseline			Baseline + Revealed		
	(1)	(2)	(3)	(4)	(5)	(6)
Subj. prob. of type A	-1.605**	-1.591**	-1.157**	-1.876***	-1.751***	-1.558***
	(0.445)	(0.533)	(0.364)	(0.398)	(0.494)	(0.481)
Purple	-0.051	-0.075	0.045	-0.267	-0.289	-0.289
	(0.412)	(0.468)	(0.518)	(0.193)	(0.202)	(0.250)
Purple × Subj. prob. of type A	0.392	0.438	0.298	0.685	0.764	0.791
	(0.868)	(0.963)	(0.884)	(0.498)	(0.536)	(0.535)
Effect of belief on purple	-1.213	-1.153	-0.859	-1.191**	-0.987	-0.766
Session FEs		✓	✓		✓	✓
Individual Controls			✓			✓
Baseline mean	2.750	2.750	2.750	2.968	2.968	2.968
Observations	96	96	96	192	192	192
R^2	0.227	0.231	0.317	0.126	0.186	0.236

Notes: Linear fixed effects model estimation of the effect of the belief of employer being type A, the taste-based discriminator, on respondent's fairness views of the hiring procedure after round 4. All specifications include controls for the cumulative average share of hires that are purple by round 4. Specifications 4-6 include a treatment dummy. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Baseline mean is the average reported fairness of the included sample from an average subjective probability of the employer being type A of 44.4% (columns 1-3) and 22.2% (columns 4-6) where the subjective probability is hard-coded to be 0% for the *revealed* treatment participants. Standard errors clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

treatment, I include observations from this arm in specifications 4 through 6. In this case, the effect of beliefs on green subjects on the perceived fairness of the labour market remains negative and strongly significant whereas the effect is not distinguishable from zero for purple participants when including session fixed effects and individual controls. This result is robust to using the same elicitation at the end of the experiment (see Table E.6).

Result 3: The more probable an individual from a discriminated group believes the outcome is due to taste-based discrimination, the more unfair they perceive the labour market, conditional on a level of discrimination observed.

Temporary affirmative action demand and perceived efficacy. I investigated if individual's willingness to pay for one-round of affirmative action was impacted by their perceptions of the employer's type. I explained to participants that in round five, an affirmative action policy in which hiring decisions would be ex-ante colour blind may be implemented. Theoretically, the policy fixes a belief that an unbiased employer would set equal signal thresholds for each group, and therefore may assist in re-coordinating the labour market out of discrimination traps once it is lifted in the fol-

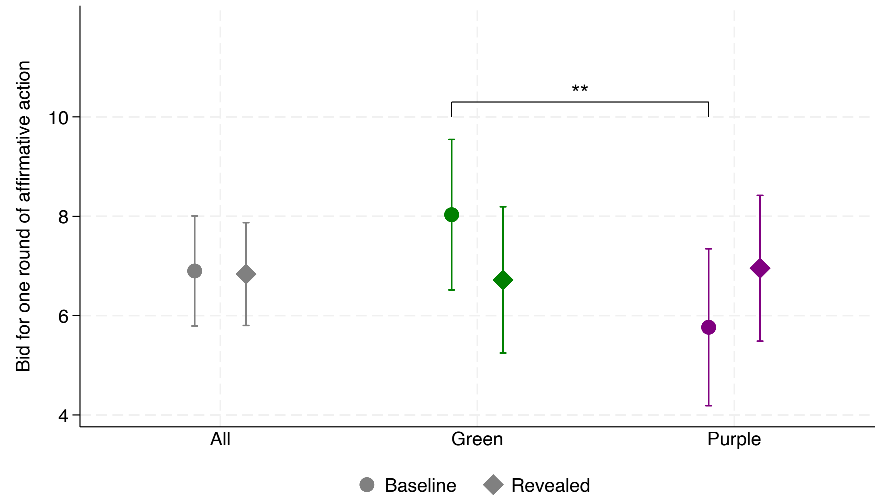


Figure 8: Demand for temporary affirmative action policy by colour and treatment

Note: This graph shows the average bid for one period of affirmative policy to be enacted in round 5. Participants could bid between 0 and 20 tokens and place their bid after being informed of the true employer type, if in *revealed* treatment. Error bars show 95% confidence intervals. Stars show the significance of the difference computed from a two-sided t-test: *p<0.10, **p<0.05, ***p<0.01.

lowing rounds. However, this intuition relies on the perceived bias of the employer: any prejudice would restrict the degree to which signal thresholds would converge. If the employer is believed to be a taste-based discriminator, one would believe that the benefits of such policy will be short-lived and the economy would revert to the non-discriminatory equilibrium afterwards. However, if the employer is perceived to be a statistical discriminator, then the benefits may outlast the policy: re-coordinating application strategies to the pooling equilibrium will erase any discrimination in the long-run.

The willingness to pay elicitation followed a modified Becker-DeGroot-Marschak mechanism (1964). Participants could bid up to 20 tokens for the policy to be enacted. One participant in a society was randomly chosen as the decision maker and a random number between 0 and 20 was drawn. If the decision maker placed a bid below the random number then the affirmative action policy was not enacted. If the decision maker's bid was at or above the random number then the policy was enacted in round five and the decision maker was deducted a number of tokens equal to the random number.

In each of the baseline and *revealed* treatment arms, the quota was enacted in one society. As shown in Figure 8, there was no difference in average bids for the quota between the baseline and *revealed* treatments (6.898 vs 6.835 respectively, two-sided t-test $t = 0.081$ $p = 0.936$). Looking at each identity individually, informing participants that the employer is not a taste-based discriminator reduces the average bid among the discriminated group and raises it among the favoured group, although neither difference is statistically significant. While there is no difference between the average bids of green and purple participants in the *revealed* treatment (6.719 (green) vs. 6.953 (purple), two-sided t-test $t = -0.221$ $p = 0.825$), the average purple bid is significantly less than the average green bid when the employer type is ambiguous (8.031 (green) vs. 5.766 (purple), two-sided t-test $t = 2.023$ $p = 0.045$). Figure E.6 shows that these results are robust to the exclusion of participants in societies who observe a hiring history that favours green applicants.

This result suggests that when the employer may be a taste-based discriminator, the discrim-

Table 4: Effect of misattribution on demand for temporary affirmative action

Dependent Variable:	Bid for one period of quota					
	Baseline			Baseline + Revealed		
	(1)	(2)	(3)	(4)	(5)	(6)
Subj. prob. of type A	2.063	1.711	1.159	1.267	0.780	0.906
	(2.822)	(3.202)	(2.498)	(2.272)	(2.563)	(2.149)
Purple	1.634	1.950	0.924	0.567	0.632	-0.021
	(3.136)	(3.169)	(2.827)	(1.163)	(1.182)	(1.003)
Purple × Subj. Prob of type A	-8.710	-9.397	-8.797	-7.058**	-7.328**	-7.095***
	(5.487)	(5.412)	(4.781)	(2.832)	(2.689)	(2.401)
Effect of belief on purple	-6.647	-7.686*	-7.638 **	-5.791**	-6.548***	-6.189***
Session FEs		✓	✓		✓	✓
Individual Controls			✓			✓
Baseline mean	6.898	6.898	6.898	6.867	6.867	6.867
Observations	128	128	128	256	256	256
R^2	0.085	0.155	0.276	0.054	0.135	0.242

Notes: Linear fixed effects model estimation of the effect of the subjective probability of employer being type A, the taste-based discriminator, on respondent's bid for one period of affirmative action. Columns 1-3 use only baseline data while columns 4-6 include the *revealed* treatment adding a dummy control for the treatment condition. All specifications include controls for the cumulative average share of hires that are purple by round 4. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

inated group places a higher value on temporary affirmative action than the favoured group. To test if this effect is robust to controlling for the level of inequality, I estimate a series of linear regression models of participants' bids on their subjective probability that the employer is type A, controlling for the level of discrimination in their society. The results are displayed in Table 4. When including session fixed effects and individual controls, I find that, conditional on a hiring history, the marginal effect of beliefs on quota demand is dependent on the assigned group identity of the participant. The more likely the employer is perceived to be type A, the lower the predicted bid among the favoured group. Conversely, the effect of the beliefs of employer type does not affect the predicted bids among the discriminated group. As the policy is only perceived to have lasting benefits if the employer is perceived to be a statistical discriminator, this can be interpreted as evidence of the self-serving bias discussed above. [28]

This result is consistent with a self-serving bias and myopic expectations on future behaviours: that participants do not internalise the long-term potential of such a quota, and are their demand is therefore only sensitive to the perceived short-term benefits or costs to their apparent favour. To

[28] In support of this interpretation, Figure E.9 demonstrates that, in the baseline treatment where the employer type is ambiguous, bids are slightly increasing in green's subjective probability of the employer being type A whereas they are decreasing for purple workers. Figure E.10 shows that this relationship does not hold as strongly when analysing the correlation between round 4's reported probability of the employer being type A on the bid for participants in the *revealed* treatment.

investigate this mechanism, I analyse participants' responses to a set of survey questions. The first survey question was asked right after placing a bid for the quota policy: participants were asked to report how likely they perceived the quota to increase the share of green hires after round 5, if it were to be enacted. Responses were recorded on a 5-point Likert scale going from 1 being Very Unlikely to 5 being Very Likely.

On average, participants in the baseline and *revealed* treatments reported very similar scores slightly leaning towards the quota being more likely to have a long-run impact than not likely (3.156 vs. 3.195, two-sided t-test $t = -0.300$, $p = 0.764$). To test for myopic expectations, I run a series of linear regressions of the reported likelihood of quota efficacy on the subjective probability that the employer is type A, controlling for the hiring history observed. The estimated coefficients displayed in Table E.8 are consistent with such expectations: conditional on a level of inequality in the society, beliefs of employer type reported in round 4 do not affect the perceived efficacy of a quota.

Table E.9 displays the results of re-estimating the regression specifications in Table 4 additionally including the ex-ante perceived efficacy of the quota and interacting this with the colour of the participant. In addition to the beliefs of the employer type, the perceived efficacy of the quota strongly influences the reported demand for such a policy. The more likely a quota is perceived effective, the higher the estimated bid for the policy, both for the discriminated and the favoured groups. That this effect is attenuated for the favoured group suggests that some self-serving bias is at play, the analysis supports the claim the myopic expectations also play a role in dampening demand.[29]

Figure E.8 plots the treatment effect of demand for the quote on the extensive and intensive margins. Although statistically insignificant, it appears that demand effects are primarily driven along the extensive margin: learning that the employer does not have a fixed bias decreases the probability that the discriminated group makes a bid and increases the probability that the favoured group does. Revealing the true employer type does not have any effect on the magnitude of the bid for either identity. I analyse the effect of beliefs of employer type on demand for the extensive and intensive margins in Tables E.10 and E.11 respectively. The estimation results in these tables support the finding that demand is only affected on the extensive margin and the effect of believed employer type on demand works in opposite directions depending on the identity of the participant.

Collecting the results in this section, demand for a potentially re-coordinating policy hinges on both the belief of the employer type and the perceived efficacy of a policy in equalising the playing field. The more effective a policy is expected to be in reducing discrimination, the higher the demand among both identities, however, the more likely the discrimination is perceived to be due to a fixed bias, demand spreads: the discriminated group places higher value on the quota while the favoured group places lower value. Therefore, lower observed demand among the favoured group compared to the discriminated group can be attributed to both self-serving motives and low expectations of success.

Result 4: Conditional on the level of discrimination, the more probable the employer is believed to be a taste-based discriminator, the lower the willingness to pay for temporary affirmative action among the favoured group.

[29]On inspecting the results of the survey at the end of the experiment. Participants in the *revealed* treatment who did not experience a quota stated that they believed the quota would have been more effective in increasing the share of green applicants and hires than those in the baseline (see Figure E.11). This effect is statistically significant at the 5% level for purple participants regarding the application rate and at the 10% level for green participants for the hiring rate. Thus, there appears to be some evidence for treatment differences in ex-post perceptions of the efficacy of affirmative action.

5 Discussion

Section 4 demonstrates that, upon observing discrimination, subjects overestimate the probability that such comes from preferences over accurate beliefs, compared to a Bayesian benchmark. Misattribution of the source of discrimination increases the likelihood that labour supply strategies in a job-application game diverge according to identity, a mechanism which sustains the underlying statistical discrimination that is misperceived. This result corresponds with the theoretical predictions of the model. In the remainder of this section, I discuss elements of the experiment which are not precisely predicted by the model: the indirect effects of misattribution on the stability of discrimination and the underlying misperceptions leading to misattribution; connecting this to policy recommendations, and highlighting the scope of these claims.

The experiment reveals two indirect channels through which misattribution increases friction in re-coordinating the market. First, participants perceived discrimination as more unfair when they believed it more likely caused by prejudice than by statistics. Since fairness views of labour markets have been shown to influence labour supply decisions, this suggests additional frictions in attracting candidates from previously under-represented groups when discrimination is misattributed (Charness and Levine, 2000). An alternative interpretation is that this result reflects a broader preference for meritocratic process, a common finding across settings (Almås et al., 2024, 2026). Either way, beliefs about the source of discrimination may feed back into labour supply: individuals may withdraw their labour if they perceive a market, or a particular employer, as unfair, which would in turn reinforce the underlying statistical discrimination (Gagnon et al., 2025; Bracha et al., 2015).

Second, misattribution reduced the perceived efficacy of and willingness to pay for temporary affirmative action, particularly among the favoured group. From a policy perspective, a temporary quota is effective precisely because it re-coordinates application strategies toward the non-discriminatory equilibrium, with lasting effects after it is lifted (Gupta, 2024). However, those misattributing the source of discrimination would perceive discrimination as an inevitability rather than a correctable coordination failure, and would expect inequality to return once the quota expires. That misattribution reduces demand for affirmative action particularly among the favoured group is indicative of self-serving preferences if future effects of the policy are not internalised (Bohren et al., 2025). As the discriminated group is less sensitive to the perceived employer type in their demand for affirmative action, it appears that participants underestimated the potential for strategic responses to this policy. Disconcertingly, since the favoured group is often the one with the majority of political capital, misattribution may reduce the scope for affirmative action enactment. In applying this result to the political discourse on affirmative action implementation, the framing of affirmative action, as a correction for market mis-coordination or as a compensation for entrenched bias, may be as consequential as its design.

To generate the specific self-confirming structure of the model, I require negative selection of applicants and the initial opacity of productivity. However, the core mechanism, that misattribution eliminates perceived scope for re-coordination, does not necessarily depend on these features. The conditions required are that multiple equilibria exist under correct attribution, that only the discriminatory equilibrium survives misattribution, and that the two attributions produce observationally equivalent outcomes for agents. Thus, other models of self-confirming discrimination in, for instance, human capital investment also satisfy these criteria. Extending the mechanism to settings with performance-based pay or multiple competing employers is the subject of my ongoing work.

Nonetheless, the external validity of my results is limited by three noteworthy factors. First, the model treats misattribution as static. Although the experiment suggests that it is self-reinforcing and consistent with a path-dependent belief dynamic, a fully dynamic model would formalise this.

Second, the laboratory experiment is beneficial for internal validity as I can control for the true employer type, however this is not straightforward in other environments. For instance, an identification challenge using observational data is that traditional approaches cannot convincingly disentangle taste-based from statistical discrimination (Guryan and Charles, 2013; Canay et al., 2024; Bohren et al., 2025). In addition, social desirability bias may thwart the accurate measurement of participant type if it is perceived undesirable as to portray taste-based discrimination traits (Bursztyn et al., 2025; Listo et al., 2025). Finally, my framework does not intend to model all corners of the labour market. In cases where wages are effort-dependent or outside options carry inherent risk, the required negative selection may not be valid.

A natural line of inquiry following these results asks what underlying misperceptions lead to the misattribution of discrimination source. In the theoretical model, I propose that cursedness, the misperception of other agents' strategies, is a candidate cause of misattribution. In Appendix C, I report results from an experimental treatment arm which alleviates participants of the strategic uncertainty by informing them of the application strategies of others in their society. I show that subjects did not systematically believe in pooling application strategies nor did the treatment alleviating strategic uncertainty relieve them of their attribution bias. This leaves an open question as to why participants misperceived the source of discrimination. Enke (2020) suggests that this may be a matter of complexity: agents may misperceive the mapping of separating application strategies to rational discrimination. This is compelling since a large portion of my sample who misattribute the source state they form their beliefs looking solely at the hiring histories, even though they correctly perceive separating application strategies. If true then, participants may attribute discrimination to taste-based sources as this justification offers the path of least mental resistance. Another explanation could be signal neglect (Giglio and Shue, 2014; Araujo et al., 2021; Graeber, 2023; Barron et al., 2024). Individuals may not account for the fact that they do not observe the true qualities of applicants and thus the possible distribution of hiring histories from which they may observe. Finally, individuals may employ representativeness heuristics in updating their beliefs on the probable cause of discriminatory outcomes (Bordalo et al., 2021, 2026). In doing so, individuals see that discrimination is more representative of the taste-based discriminator and so assign a higher probability to that being the cause of the inequality. I leave this investigation to that of future research.

6 Conclusion

A number of authors have demonstrated the theoretical existence of self-confirming discrimination across various domains. In this equilibrium, discrimination represents a coordination error. However, little is known about how such coordination errors arise and persist. This paper argues that the answer lies just not in whether people anticipated about discrimination, but also in what they believe causes it: misattributing statistical to taste-based discrimination leads to stickier inequality.

The present theoretical model applies the self-confirming discrimination framework of Coate and Loury (1993) to the job search model of Lucas (1978). When workers correctly attribute the discrimination to an informationally constrained unbiased employer, they foresee the outcome as a coordination failure: they could reach a non-discriminatory equilibrium with an organised shift in strategies. When workers instead misattribute the source of discrimination to prejudicial preferences, discrimination becomes stickier. The non-discriminatory equilibrium is perceived as attainable and so re-coordination is undervalued. Moreover, the outcomes do not challenge beliefs and so misattribution can be stable.

Results from a pre-registered laboratory experiment supports both the assumptions and the predictions of this model. Facing a discriminatory hiring history, subjects assign a 57% probability to a taste-based employer, significantly above the 50% Bayesian benchmark, and this error is not benign: a higher subjective probability of taste-based discrimination lowers predicted application rates for the discriminated group and raises them for the favoured group. This effect concentrated exactly where the model predicts it should be sharpest, among workers on the margin of applying. Moreover, I find that subjects who exhibit misattribution express an increased sense of market unfairness when in the discriminated group and a lower willingness to pay for an affirmative action policy when in the favoured group. Both of these effects may have profound consequences in the ability to re-coordinate the market to an outcome presenting less discrimination.

To conclude, the evidence presented in this paper paints a compelling story explaining the emergence and persistence of discrimination traps. As the true prevalence of prejudiced employers has fallen, the belief in them has not kept pace, and that misperception alone is enough to sustain inequality, distort labour supply, and block the policies that would restore coordination. That gap between reality and belief, this paper argues, is where discrimination hides.

References

Abeler, J., Altmann, S., Kube, S., and Wibral, M. (2010). Gift exchange and workers' fairness concerns: when equality is unfair. *Journal of the European Economic Association*, 8(6):1299–1324.

Agüero, J. M., Galarza, F., and Yamada, G. (2023). (Incorrect) perceived returns and strategic behavior among talented low-income college graduates. *AEA Papers and Proceedings*, 113:423–426.

Aksoy, B., Chadd, I., and Koh, B. H. (2023). Sexual identity, gender, and anticipated discrimination in prosocial behavior. *European Economic Review*, 154:104427.

Almås, I., Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2024). Attitudes to inequality: preferences and beliefs. *Oxford Open Economics*, 3(Supplement_1):i64–i79.

Almås, I., Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2026). Fairness across the world. *Quarterly Journal of Economics*.

Alston, M. (2025). The (perceived) cost of being female: An experimental investigation of strategic responses to discrimination. Working paper.

Angeli, D., Matavelli, I., and Secco, F. (2023). Expected discrimination and job search. Working paper.

Araujo, F. A., Wang, S. W., and Wilson, A. J. (2021). The times they are a-changing: Experimenting with dynamic adverse selection. *American Economic Journal: Microeconomics*, 13(4):1–22.

Arrow, K. J. (1974). The theory of discrimination. In *The Limits of Organization*, pages 55–82. W.W. Norton & Company, New York.

Bandiera, O., Jalal, A., and Roussille, N. (2025). The illusion of time: Gender gaps in job search and employment. Working Paper (No. w34051), National Bureau of Economic Research.

Barron, K., Huck, S., and Jehiel, P. (2024). Everyday econometricians: Selection neglect and overoptimism when learning from others. *American Economic Journal: Microeconomics*, 16(3):162–198.

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232.

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

Benjamin, D. J., Moore, D. A., and Rabin, M. (2017). Biased beliefs about random samples: Evidence from two integrated experiments. Working Paper (No. w23927), National Bureau of Economic Research.

Bertrand, M. and Duflo, E. (2017). Field experiments on discrimination. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Economic Field Experiments*, volume 2, pages 309–393. North-Holland.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2025). Inaccurate statistical discrimination: An identification problem. *Review of Economics and Statistics*, pages 1–16.

Bohren, J. A. and Hauser, D. N. (2021). Learning with heterogeneous misspecified models: Characterization and robustness. *Econometrica*, 89(6):3025–3077.

Bohren, J. A. and Hauser, D. N. (2025). Misspecified models in learning and games. *Annual Review of Economics*, 17.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109(10):3395–3436.

Bordalo, P., Coffman, K., Gennaioli, N., Schwerter, F., and Shleifer, A. (2021). Memory and representativeness. *Psychological Review*, 128(1):71.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *Quarterly Journal of Economics*, 131(4):1753–1794.

Bordalo, P., Conlon, J., Gennaioli, N., Kwon, S., and Shleifer, A. (2026). How people use statistics. *Review of Economic Studies*, 93(1):250–285.

Boring, A., Coffman, K., Glover, D., and González-Fuentes, M. J. (2025). Discrimination, rejection, and job search. Working paper.

Bracha, A., Gneezy, U., and Loewenstein, G. (2015). Relative pay and labor supply. *Journal of Labor Economics*, 33(2):297–315.

Breza, E., Kaur, S., and Shamdasani, Y. (2018). The morale effects of pay inequality. *Quarterly Journal of Economics*, 133(2):611–663.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Bursztyn, L., Haaland, I. K., Röver, N., and Roth, C. (2025). The social desirability atlas. Working Paper (No. w33920), National Bureau of Economic Research.

Campos-Mercade, P. and Mengel, F. (2024). Non-bayesian statistical discrimination. *Management Science*, 70(4):2549–2567.

Canay, I. A., Mogstad, M., and Mountjoy, J. (2024). On the use of outcome tests for detecting bias in decision making. *Review of Economic Studies*, 91(4):2135–2167.

Card, D., Mas, A., Moretti, E., and Saez, E. (2012). Inequality at work: The effect of peer salaries on job satisfaction. *American Economic Review*, 102(4):2981–3003.

Charles, K. K. and Guryan, J. (2008). Prejudice and wages: an empirical assessment of Beckers The Economics of Discrimination. *Journal of Political Economy*, 116(5):773–809.

Charness, G., Cobo-Reyes, R., Meraglia, S., and Sánchez, Á. (2020). Anticipated discrimination, choices, and performance: Experimental evidence. *European Economic Review*, 127:103473.

Charness, G. and Levine, D. I. (2000). When are layoffs acceptable? Evidence from a quasi-experiment. *ILR Review*, 53(3):381–400.

Chauvin, K. (2018). A misattribution theory of discrimination. Working paper.

Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree: An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Coate, S. and Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review*, pages 1220–1240.

Craig, A. C. and Fryer, R. G. (2017). Complementary bias: A model of two-sided statistical discrimination. Working Paper (No. w23811), National Bureau of Economic Research.

Dianat, A., Echenique, F., and Yariv, L. (2022). Statistical discrimination and affirmative action in the lab. *Games and Economic Behavior*, 132:41–58.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., and Davies, P. G. (2004). Seeing black: race, crime, and visual processing. *Journal of Personality and Social Psychology*, 87(6):876.

Enke, B. (2020). What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398.

Enke, B. and Graeber, T. (2023). Cognitive uncertainty. *Quarterly Journal of Economics*, 138(4):2021–2067.

Enke, B., Graeber, T., and Oprea, R. (2025). Complexity and time. *Journal of the European Economic Association*, 23(5):1838–1867.

Enke, B., Graeber, T., Oprea, R., and Yang, J. (2024). Behavioral attenuation. Working Paper (No. w32973), National Bureau of Economic Research.

Esponda, I., Oprea, R., and Yuksel, S. (2023). Seeing what is representative. *Quarterly Journal of Economics*, 138(4):2607–2657.

Esponda, I., Pouzo, D., and Yamamoto, Y. (2021). Asymptotic behavior of bayesian learners with misspecified models. *Journal of Economic Theory*, 195:105260.

Exley, C. L., Fisman, R., Kessler, J. B., Lepage, L.-P., Li, X., Low, C., Shan, X., Toma, M., and Zafar, B. (2024). Information-optional policies and the gender concealment gap. Working Paper (No. 32810), National Bureau of Economic Research.

Eyster, E. and Rabin, M. (2005). Cursed equilibrium. *Econometrica*, 73(5):1623–1672.

Eyting, M. (2022). Why do we discriminate? the role of motivated reasoning. Working Paper (No. 356), Leibniz Institute for Financial Research SAFE.

Fang, H. and Norman, P. (2006). Government-mandated discriminatory policies: Theory and evidence. *International Economic Review*, 47(2):361–389.

Filippin, A. (2009). Can workers' expectations account for the persistence of discrimination? Working Paper (No. 4490), IZA discussion papers.

Flory, J. A., Leibbrandt, A., and List, J. A. (2018). Signaling interest in diversity backfires for racial minorities. Working Paper (No. 25035), National Bureau of Economic Research.

Fluchtmann, J., Glenny, A. M., Harmon, N. A., and Maibom, J. (2024). The gender application gap: Do men and women apply for the same jobs? *American Economic Journal: Economic Policy*, 16(2):182–219.

Frick, M., Iijima, R., and Ishii, Y. (2023). Belief convergence under misspecified learning: A martingale approach. *The Review of Economic Studies*, 90(2):781–814.

Frydman, C. and Jin, L. J. (2022). Efficient coding and risky choice. *Quarterly Journal of Economics*, 137(1):161–213.

Fudenberg, D., Lanzani, G., and Strack, P. (2021). Limit points of endogenous misspecified learning. *Econometrica*, 89(3):1065–1098.

Gabaix, X. and Graeber, T. (2024). The complexity of economic decisions. Working Paper (No. 33109), National Bureau of Economic Research.

Gagnon, N., Bosmans, K., and Riedl, A. (2025). The effect of gender discrimination on labor supply. *Journal of Political Economy*, 133(3):000–000.

Gallup (2025). Race relations. `https://news.gallup.com/poll/1687/Race-Relations.aspx`. Accessed: 2025-07-11.

Gennaioli, N. and Shleifer, A. (2010). What comes to mind. *Quarterly Journal of Economics*, 125(4):1399–1433.

Gerdeman, D. (2017). Minorities who whitenjob resumes get more interviews. *HBS Working Knowledge*, 17.

Giglio, S. and Shue, K. (2014). No news is news: do markets underreact to nothing? *The Review of Financial Studies*, 27(12):3389–3440.

Glover, D., Pallais, A., and Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *Quarterly Journal of Economics*, 132(3):1219–1260.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, 112(2):631–645.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119.

Graeber, T. (2023). Inattentive inference. *Journal of the European Economic Association*, 21(2):560–592.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125.

Gupta, N. (2024). Can temporary affirmative action improve representation? *Available at SSRN 4890176*.

Guryan, J. and Charles, K. K. (2013). Taste-based or statistical discrimination: The economics of discrimination returns to its roots. *The Economic Journal*, 123(572):F417–F432.

Haggag, K., Pope, D. G., Bryant-Lees, K. B., and Bos, M. W. (2019). Attribution bias in consumer choice. *The Review of Economic Studies*, 86(5):2136–2183.

Hübert, R. and Little, A. T. (2023). A behavioural theory of discrimination in policing. *The Economic Journal*, 133(655):2828–2843.

Kudashvili, N. and Lergetporer, P. (2022). Minorities' strategic response to discrimination: Experimental evidence. *Journal of Public Economics*, 208:104630.

Kuhn, P., Osaki, T., and Yue, L. (forthcoming). When is discrimination unfair? *Journal of Labor Economics*.

Lang, K. and Manove, M. (2011). Education and labor market discrimination. *American Economic Review*, 101(4):1463–1496.

Leibbrandt, A. and List, J. A. (2018). Do equal employment opportunity statements backfire? evidence from a natural field experiment on job-entry decisions. *Management Science*, 66(5):1807–1827.

Lepage, L.-P. (2024). Experience-based discrimination. *American Economic Journal: Applied Economics*, 16(4):288–321.

Lepage, L.-P., Li, X., and Zafar, B. (2025). Anticipated discrimination and major choice. Working Paper (No. 33680), National Bureau of Economic Research.

Listo, A., Muñoz, E., and Sansone, D. (2025). Measuring the sources of taste-based discrimination using list experiments. Working paper.

Lucas, R. E. (1978). On the size distribution of business firms. *Bell Journal of Economics*, pages 508–523.

Martínez-Marquina, A., Niederle, M., and Vespa, E. (2019). Failures in contingent reasoning: The role of uncertainty. *American Economic Review*, 109(10):3437–3474.

Moro, A. and Norman, P. (2003). Affirmative action in a competitive economy. *Journal of Public Economics*, 87(3-4):567–594.

Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, 46(8):1423–1437.

Onuchic, P. (2022). Recent contributions to theories of discrimination. *arXiv preprint arXiv:2205.05994*.

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology*, 108(5):937–975.

Peterson, C. R., DuCharme, W. M., and Edwards, W. (1968). Sampling distributions and probability revisions. *Journal of Experimental Psychology*, 76(2p1):236.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review*,

62(4):659–661.
Poschke, M. (2013). Who becomes an entrepreneur? labor market prospects and occupational choice. *Journal of Economic Dynamics and Control*, 37(3):693–710.
Rabin, M. and Vayanos, D. (2010). The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, 77(2):730–778.
Rackstraw, E. (2022). Bias-motivated updating in the labor market. *Available at SSRN 4278076*.
Ruebeck, H. (2024). Causes and consequences of perceived workplace discrimination. *Available at SSRN 4799864*. Working Paper.
Ruzzier, C. A. and Woo, M. D. (2023). Discrimination with inaccurate beliefs and confirmation bias. *Journal of Economic Behavior & Organization*, 210:379–390.
Samek, A. (2019). Gender differences in job entry decisions: A university-wide field experiment. *Management Science*, 65(7):3272–3281.
Sarsons, H. (2017). Interpreting signals in the labor market: evidence from medical referrals. Working paper.
Steiner, J. and Stewart, C. (2016). Perceiving prospects properly. *American Economic Review*, 106(7):1601–1631.
Teigen, K. H. (1974). Subjective sampling distributions and the additivity of estimates. *Scandinavian Journal of Psychology*, 15(1):50–55.
Ugalde, M. P. (2022). Gender, grade sensitivity, and major choice. Working paper.
Wheeler, G. and Beach, L. R. (1968). Subjective sampling distributions and conservatism. *Organizational Behavior and Human Performance*, 3(1):36–46.
Zhang, H. and Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience*, 6:1.

A Survey

In this section, I establish that individuals systemically misperceive the degree to which prejudicial preferences are distributed today. To measure such misperceptions directly, I conducted a survey in September 2025 using a sample of 401 participants based in the United States. The sample was representative of age, gender, ethnicity and political affiliation. Participants were asked two questions on whether they personally (i) support marriage between black and white people, and (ii) would vote for a well-qualified female presidential candidate from their own party. These questions were chosen since they are well-established in the Gallup World Poll and closely resemble dimensions of taste-based discrimination. Respondents then stated their belief about the share of the other 400 respondents who responded negatively to these questions.[30]

Responses to the personal views closely match the latest results from the Gallup World Poll (2025). As presented in Table A.5, 4.0% of my respondents stated that they disapproved of interracial marriage and 5.5% of my sample would not vote for the female candidate. For both questions, respondents overestimated the share of others responding negatively by almost 500%. For interracial marriage, the believed share disapproving was 22.2% while for female presidential candidate support, the believed share stating they would not vote was 30.2%. Both of these are significantly greater than the true values (two-sided t-tests $p < 0.001$).

Even though private preferences have improved in accepting and supporting minorities, societal views of the preferences of the average individual are sticky. In the next section, I present a theoretical model which establishes how such sticky preferences may sustain discrimination and thus inequality.

Table A.5: Overestimation of prevalence of prejudicial preferences

	Share responding negatively	Believed share responding negatively	T-test mean difference
	(1)	(2)	(3)
Q1: Interracial Marriage	4.0 (0.010)	22.2 (0.010)	-18.2*** (0.013)
Q2: Vote for Female Candidate	5.5 (0.011)	30.2 (0.010)	-24.8*** (0.014)

Notes: Survey questions asked to a representative sample of 401 US participants on Prolific. Representation was computed according to age, gender, ethnicity and political affiliation of the general population. Question 1 asked *Do you approve or disapprove of marriage between black people and white people?* and question 2 asked *Would you vote for a well-qualified presidential candidate from your own political party if that person happened to be a female?* Standard errors are presented in parentheses. Stars from two-sided t-test * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

[30] These beliefs were incentivised: for each question, participants who stated a belief closest to the true value were entered into a draw to win \$20. This was on top of the \$0.68 each respondent received for participating. The survey was coded in oTree (Chen et al., 2016) and respondents were recruited using Prolific.

B Proofs

Lemma 1

An individual i from group g with ability q_i is indifferent to applying to the job when

$$q_i = \frac{\omega}{2}\left(1 + \text{erf}\left(\sqrt{\frac{\eta}{2}}(q_i - \underline{\hat{s}}_g)\right)\right) \equiv y(q_i; \underline{\hat{s}}_g). \tag{9}$$

Define $\imath(q) = y(q; \underline{\hat{s}}_g) - q$. As $q \to -\infty$, $\text{erf}(\sqrt{\eta/2}(q - \underline{\hat{s}}_g)) \to -1$, so $y \to 0$, and thus $\imath(q) = y - q \to +\infty$.[31] As $q \to +\infty$, $y \to \omega$, so $\imath(q) \to \omega - \infty < 0$. As $\imath(q)$ is continuous on q, the intermediate value theorem guarantees at least one root i.e. one fixed point of y.

For uniqueness, note that the derivative of y with respect to q is

$$\frac{\partial y}{\partial q} = \omega\sqrt{\frac{\eta}{2\pi}}e^{-\frac{\eta}{2}(q-\underline{\hat{s}}_g)^2}. \tag{10}$$

This is maximised at $q = \underline{\hat{s}}_g$, where the RHS equals $\omega\sqrt{\eta/2\pi}$. Under Assumption 1, $\omega \leq \sqrt{2\pi/\eta}$, so $\partial y/\partial q \leq 1$ for all $q \in \mathbb{R}$. Therefore $\imath'(q) = \partial y/\partial q - 1 \leq 0$ for all q with strict inequality except possibly at the bound if $q = \underline{\hat{s}}_g$. Since $\imath'(q) = 0$ holds at most at the single point $q = \underline{\hat{s}}_g$ and $\imath'(q) < 0$ everywhere else, $\imath$ is strictly decreasing on $\mathbb{R}$ and crosses zero exactly once.■

Corollary 1

By Lemma 1, the unique fixed point $\bar{q}_g$ satisfies $\bar{q}_g = y(\bar{q}_g; \underline{\hat{s}}_g)$. Since the range of $y(,\cdot,; \underline{\hat{s}}_g)$ is $(0, \omega)$, it follows immediately that $\bar{q}_g \in (0, \omega)$. ■

Lemma 2

Suppose for contradiction that the employer believes $\hat{\hat{q}}_A = \hat{\hat{q}}_B = \hat{\hat{q}}$ and sets, without loss of generality, $\underline{s}_A > \underline{s}_B$. Since both groups are believed to play the same cut-off strategy $\hat{\hat{q}}$, the distribution of productivity conditional on application is identical across the two groups. By equation (7), the posterior expected productivity of an applicant from group g with signal s is

$$\mathbb{E}[q|s; q \leq \hat{\hat{q}}] = \frac{t(\hat{\hat{q}}) \cdot u(\hat{\hat{q}}) + \eta s}{t(\hat{\hat{q}}) + \eta}$$

which does not depend on g. For any signal $s \in [\underline{s}_B, \underline{s}_A)$, a group B applicant is hired while a group A applicant with the same signal is rejected, yet both yield identical expected profit to the employer. Since expected profit is strictly increasing in s, the employer would strictly prefer to replace any hired group B worker at signal $s \in [\underline{s}_B, \underline{s}_A)$ with a rejected group A applicant yielding a slightly higher signal $s' \in (\underline{s}_B, \underline{s}_A)$.

Raising $\underline{s}_B$ by any $\epsilon_B > 0$ while decreasing $\underline{s}_A$ by some $\epsilon_A > 0$ such that the expected profits from hiring are unchanged, and continuing this until $\underline{s}_A = \underline{s}_B$ is weakly preferred. Therefore, any $\underline{s}_A = \underline{s}_B$ dominates $\underline{s}_A \neq \underline{s}_B$, contradicting the optimality of $\underline{s}_A > \underline{s}_B$.■

[31] The error function, erf, is used throughout the proofs. The function arises from the cumulative distribution function of the normal distribution.

Proposition 1

By Lemma 2, if an equilibrium is defined by application strategies $\bar{q}_A = \bar{q}_B = \bar{q}$, it must be defined by hiring thresholds of $\underline{s}_A = \underline{s}_B = \underline{s}(\bar{q})$.

Likewise, by Lemma 1, for any equilibrium $\underline{s}(\bar{q})$, there exists a unique equilibrium cut-off $\bar{q}$ which satisfies

$$\underline{s}(\bar{q}) = \bar{q} - \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(\frac{2\bar{q}}{\omega} - 1\right). \tag{11}$$

This is well-defined for $\bar{q} \in (0, \omega)$ since erf^{-1} has domain $(-1, 1)$, confirmed by Corollary 1.

Optimality requires that the following equation holds,

$$\mathbb{E}[q|\underline{s}(\bar{q}_A); q \leq \bar{q}_A] \equiv E(\bar{q}_A) = E(\bar{q}_B) \equiv \mathbb{E}[q|\underline{s}(\bar{q}_B); q \leq \bar{q}_B]. \tag{12}$$

which is trivially satisfied for any $\bar{q}_A = \bar{q}_B$ in $(0, \omega)$. Thus, if the non-discriminatory equilibrium exists and is unique, then it must be defined by the satiation equilibrium condition.

For a pooling equilibrium defined by $\bar{q}$ to satisfy the satiation condition (6), it must be that

$$\phi = \sum_g n_g \beta^g(h(\phi, \psi, m, \bar{q}), \bar{q}) \equiv \sum_g n_g \int_{-\infty}^{\bar{q}} Pr\left(e \geq \underline{s}(\bar{q}) - q|q\right) f(q)dq, \tag{13}$$

where $e \sim N(0, 1/\eta)$ is a white noise term of the signal. Setting the integral in (13) as $R(\bar{q})$, the above condition becomes $\phi = R(\bar{q})$. Substituting for $\underline{s}(\bar{q})$, we can write $R(\bar{q})$ as,

$$R(\bar{q}) = \int_{-\infty}^{\bar{q}} Pr\left(e \geq \bar{q} - q - \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(\frac{2\bar{q}}{\omega} - 1\right)|q\right) f(q)dq. \tag{14}$$

Now I consider the boundary values of $R(\bar{q})$:

1. At $\bar{q} \to 0^+$: Since $\text{erf}^{-1}(-1) = -\infty$, the threshold satisfies $\underline{s}(\bar{q}) \to +\infty$. This means that $Pr(\epsilon \geq \underline{s}(\bar{q}) - q|q) \to 0$ for all finite q as such $R(\bar{q}) \to 0$.

2. At $\bar{q} \to \omega^-$:Since $\text{erf}^{-1}(1) = +\infty$, the threshold satisfies $\underline{s}(\bar{q}) \to -\infty$. Therefore $Pr(\epsilon \geq \underline{s}(\bar{q}) - q|q) \to 1$ for all q, and $R(\bar{q}) \to \int_{-\infty}^{\omega} f(q)dq = F(\omega)$.

Applying the Leibniz rule to differentiate $R(\bar{q})$,

$$R'(\bar{q}) = Pr(e \geq \underline{s}(\bar{q}) - q|q = \bar{q})f(\bar{q}) + \int_{-\infty}^{\bar{q}} \frac{\partial}{\partial \bar{q}} Pr\left(e \geq \bar{q} - q - \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(\frac{2\bar{q}}{\omega} - 1\right)|q\right) f(q)dq. \tag{15}$$

The first term is strictly positive since $f_q(\bar{q}) > 0$, probabilities are non-negative and q has full support. For the second term, let $\psi_e(x)$ be the density of e, then

$$\frac{\partial}{\partial \bar{q}} Pr\left(e \geq \bar{q} - q - \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(\frac{2\bar{q}}{\omega} - 1\right)|q\right) = -\psi_e(\underline{s}(\bar{q}) - q)\underline{s}'(\bar{q}).$$

As a density is always strictly positive, the remaining fact is to show that $\underline{s}'(\bar{q})$ is strictly negative. To do this, I rely on Claim I.

Claim I: Given assumption 1, the equilibrium signal threshold is weakly decreasing in the application cutoff value, $\underline{s}'(\bar{q}) \leq 0$.

Proof: Using (11), the derivative of the optimal signal threshold is,

$$\underline{s}'(\bar{q}) = 1 - \frac{1}{\omega}\sqrt{\frac{2\pi}{\eta}} e^{[\text{erf}^{-1}(\frac{2\bar{q}}{\omega} - 1)]^2}.$$

By assumption 1, $\omega \leq \sqrt{2\pi/\eta}$. Therefore, since the exponential factor is at least 1, it follows that $\underline{s}'(\bar{q}) \leq 0$.■

Using Claim I and that $f(q) > 0$, it must be that the integrand in the second term on the right hand side of (15) is non-negative. Therefore it must be that $R'(\bar{q}) > 0$ for all $\bar{q} \in (0, \omega)$. Given this, $R(\bar{q})$ is monotonic and therefore any value of ϕ that satisfies assumption 2, must define a unique application strategy that constitutes a non-discriminatory equilibrium. ■

Proposition 2

Denote the signal threshold that, given application strategy $\bar{q}_g$, satisfies (4) as $\underline{s}(\bar{q}_g)$. To satisfy the optimality condition of Definition 1, application strategies and signal thresholds must satisfy,

$$\mathbb{E}[q|\underline{s}(\bar{q}_A); q \leq \bar{q}_A] \equiv E(\bar{q}_A) = E(\bar{q}_B) \equiv \mathbb{E}[q|\underline{s}(\bar{q}_B); q \leq \bar{q}_B]. \tag{16}$$

From Lemma 1, for every signal threshold, there is a unique application strategy that will satisfy (4). Substituting (11) into (16), I can write the new condition as,

$$\frac{t(\bar{q}_A)u(\bar{q}_A) + \eta\underline{s}(\bar{q}_A)}{t(\bar{q}_A) + \eta} = \frac{t(\bar{q}_B)u(\bar{q}_B) + \eta\underline{s}(\bar{q}_B)}{t(\bar{q}_B) + \eta}. \tag{17}$$

Clearly, the above is satisfied when $\bar{q}_A = \bar{q}_B$, which permits the existence of the non-discriminatory equilibrium outlined in Proposition 1. The current proposition proposes that there are some values of ω and ϕ, such that (17) and (6) may be jointly satisfied when $\bar{q}_A \neq \bar{q}_B$. To show this, I require the additional lemma.

Lemma 3: Given that Assumption 1 holds, there exists some non-empty set Ω such that if $\omega \in \Omega$, then there is some non-empty set $\tilde{Q} \subset (0, \omega)$ such that if $\bar{q} \in \tilde{Q}$, $E'(\bar{q}) > 0$.
Proof: I begin with 2 further claims that I will refer to in the proceeding proof.

Claim II: Let the mean of the normal distribution $N(\mu, 1/\tau)$ truncated from above at a, be denoted by $u(a)$;

- $u'(a) = \theta(z(a))[z(a)+\theta(z(a))]$ where $\theta(z(a))$ is the inverse Mill's ratio and $z(a) = \sqrt{\tau}(a-\mu)$ is the standard score.
- $u'(a) > 0$.

Proof: The described mean $u(a)$ may be written as

$$u(a) = \mu - \frac{\theta(z(a))}{\sqrt{\tau}} \tag{18}$$

where

$$z(a) = \sqrt{\tau}(a - \mu),\ \theta(x) = \frac{\Psi(x)}{c(x)},\ \Psi(x) \equiv \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},\ c(x) \equiv \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right].$$

Using the functional forms of $\Psi(\cdot)$ and $c(\cdot)$, I can find their derivatives $\Psi'(x) = -x\Psi(x)$ and $c'(x) = \Psi(x)$. Therefore, I can write $\theta'(x) = -x\theta(x) - \theta(x)^2$. Differentiating $u(a)$ with respect to a gets

$$\frac{du(a)}{da} = -\frac{\theta'(z(a))z'(a)}{\sqrt{\tau}} = -\theta'(z(a)) = \theta(z(a))\,[z(a) + \theta(z(a))]$$

Since the mean of a distributed truncated above at a is strictly less than a, it must be that $\theta(z(a)) > -z(a)$. Thus, the term inside the brackets above must be greater than zero and as such $u'(a) > 0$ for all a. ■

Claim III: The concentration of the normal distribution $N(\mu, 1/\tau)$ truncated from above at a, $t(a)$, has a gradient which can be written as the following,

$$t'(a) = t(a)\sqrt{\tau}\theta(z(a))\left[1 - \frac{(\theta + z)^2}{(1 - \theta(z(a))[\theta(z(a)) + z(a)])^2}\right], \tag{19}$$

where $\theta(z(a))$ is the inverse Mill's ratio and $z(a) = \sqrt{\tau}(a - \mu)$ is the standard score.
Proof: The concentration of the normal distribution $N(\mu, 1/\tau)$ truncated from above at a is

$$t(a) = \frac{\tau}{1 - \theta(z(a))^2 - z(a)\theta(z(a))}. \tag{20}$$

The derivative of this with respect to the truncation point can be written as,

$$t'(a) = \frac{-\tau\left[-2\theta(z(a))\theta'(z(a))z'(a) - z'(a)\theta(z(a)) - z(a)z'(a)\theta'(z(a))\right]}{(1 - \theta(z(a))^2 - z(a)\theta(z(a)))^2}$$

Substituting in for $z'(a)$ and $\theta'(z(a))$ as derived in Claim II, we can write this as the following,

$$t'(a) = \tau\sqrt{\tau}\theta(z(a))\frac{1 - [2\theta(z(a)) + z(a)][z(a) + \theta(z(a))]}{(1 - \theta(z(a))^2 - z(a)\theta(z(a)))^2}.$$

Substituting $t(a)$ back in and expanding the numerator of the remaining fraction, this gives,

$$t'(a) = t(a)\sqrt{\tau}\theta(z(a))\left[1 - \frac{(\theta(z(a)) + z(a))^2}{1 - \theta(z(a))(\theta(z(a)) + z(a))}\right] \tag{21}$$

which is the desired result.■

From corollary 1, the range of $\bar{q}_g$ than can constitute an equilibrium application strategy is $(0, \omega)$. As the lower bound of this, the equilibrium signal threshold is equal to,

$$\underline{s}(0) = 0 - \sqrt{\frac{2}{\eta}}\operatorname{erf}^{-1}(-1) = \infty.$$

At the upper bound of this range, it is equal to.

$$\underline{s}(\omega) = \omega - \sqrt{\frac{2}{\eta}}\operatorname{erf}^{-1}(1) = -\infty.$$

Therefore, $E(0) = \infty$ and $E(\omega) = -\infty$. Accordingly, $E(\bar{q})$ must be a decreasing function for some $\bar{q} \in (0, \omega)$. Using the above claims, I will now show that for some values of ω which satisfy Assumption 1, there exists a subset $\tilde{Q} \subset (0, \omega)$ such that $E'(\bar{q}) > 0$.

The first order derivative of $E(\bar{q})$ can be written as the following,

$$E'(\bar{q}) = \frac{\left[t'(\bar{q})u(\bar{q}) + u'(\bar{q})t(\bar{q}) + \eta\underline{s}'(\bar{q})\right](t(\bar{q}) + \eta) - t'(\bar{q})[t(\bar{q})u(\bar{q}) + \eta\underline{s}(\bar{q})]}{(t(\bar{q}) + \eta)^2}. \tag{22}$$

Trivially, the denominator is always positive and so I focus on identifying that there are some values of $\bar{q}$ such that the numerator is also positive. Collecting like terms, the numerator of (22) can be written as,

$$u'(\bar{q})t(\bar{q})(t(\bar{q})+\eta)+\eta[t'(\bar{q})(u(\bar{q})-\underline{s}(\bar{q}))+\underline{s}'(\bar{q})(t(\bar{q})+\eta)]. \tag{23}$$

Given Claim I, $\underline{s}'(\bar{q})$ is no greater than 0. Therefore, it is useful to consider a simplifying value of $\bar{q}$ at which the above expression is most likely to be positive: that which $\underline{s}'(\bar{q})$ is at its maximum.

A candidate value is $\bar{q}=\omega/2$ such that $\underline{s}'(\omega/2)=1-1/\omega\sqrt{2\pi/\eta}$ which, given Assumption 1, is no greater than 0. I consider the upper bound of ω such that $\omega=\sqrt{2\pi/\eta}$. In this case, $\underline{s}'(\omega/2)=0$. Substituting this into (23) and setting this to be greater than 0 gives,

$$u'(\bar{q})t(\bar{q})(t(\bar{q})+\eta)>\eta t'(\bar{q})[\underline{s}(\bar{q})-u(\bar{q})].$$

Using the fact that $\underline{s}(\omega/2)=\omega/2$ and that $u(\bar{q})=\mu-\theta(z(\bar{q}))/\sqrt{\tau}$, we can make the following substituting,

$$\underline{s}(\bar{q})-u(\bar{q})=\frac{z(a)+\theta(z(a))}{\sqrt{\tau}}.$$

Using Claim III, we can substitute the above with $t'(\bar{q})$ and simplify the right hand side such that the above inequality becomes,

$$u'(\bar{q})(t(\bar{q})+\eta)>\eta\theta(z(a))\left[1-\frac{(\theta(z(a))+z(a))^2}{1-\theta(z(a))(\theta(z(a))+z(a))}\right][z(\bar{q})+\theta(\bar{q})].$$

Using Claim II, we can substitute for $u'(\bar{q})$ and simplify to get,

$$t(\bar{q})+\eta>\eta\left[1-\frac{(\theta(z(a))+z(a))^2}{1-\theta(z(a))(\theta(z(a))+z(a))}\right].$$

Which gives the sufficient inequality

$$\frac{t(\bar{q})}{\eta}>-\frac{(\theta(z(a))+z(a))^2}{1-\theta(z(a))(\theta(z(a))+z(a))}. \tag{24}$$

Note that, by definition, the left hand side of the above must be positive. The denominator on the right hand side is equal to $t(\bar{q})/\tau$ which must be positive. To show the numerator of this fraction is positive, consider the expression of $u(a)=\mu-\theta(z(a))/\sqrt{\tau}$. Note that the mean of a normal distribution $N(\mu,1/\tau)$ truncated above at a real $a>-\infty$ must be less than the value of the truncation point a by definition. Accordingly, it must be that $\mu-\theta(z(a))/\sqrt{\tau}<a$. Rearranging this, we get that $\theta(z(a))+\sqrt{\tau}(a-\mu)>0$. Using $z(a)=\sqrt{\tau}(a-\mu)$, this means that $\theta(z(a))+z(a)>0$. Therefore, the numerator of the right hand side fraction in (24) must be positive.

This shows that $E'(\omega/2)>0$ when $\omega=\sqrt{2\pi/\eta}$. As $E(\bar{q})$ must be a decreasing function towards the bounds of the domain, and is continuous in $\bar{q}$, it must be that there exists some $\omega<\sqrt{2\pi/\eta}$ where the above holds and some $\bar{q}$ close to $\omega/2$ that this also holds. I describe the set of such as Ω and $\tilde{Q}$ respectively.■

Using Lemma 3, that $E(0)=\infty$ and $E(\omega)=-\infty$, and the continuity of $E(\bar{q})$ we can define $\tilde{Q}$ as an open set $\tilde{Q}=(Q_0,Q_1)$ such that when $\bar{q}<Q_0$ or $\bar{q}>Q_1$, $E'(\bar{q})<0$. Moreover, it must be that $E'(Q_0)=E'(Q_1)=0$. Thus, using the intermediate value theorem, we can say that for any value of $\gamma\in\tilde{Q}$, there exists some $\bar{q}_0<Q_0$ and $\bar{q}_1>Q_1$ such that $E(\bar{q}_0)=E(\gamma)=E(\bar{q}_1)$.

With this, we can describe two loci on the plane $(\bar{q}_A,\bar{q}_B)$ which map the set of values that mutually satisfy (16):

1. When $\omega < \sqrt{2\pi/\eta}$, the 45 degree line on the domain $(0, \omega)$.

2. When $\omega \in \Omega$, there is a continuous ovular locus that intersects the 45 degree line twice at (Q_0, Q_0) and (Q_1, Q_1).

The locus defined by (6) that maps the set of application strategies that mutually satisfy the satiation condition must be downward sloping in the domain as shown in the proof to Proposition 1. The intersection of this locus with the 45 degree line constitutes the non-discriminatory equilibrium. An intersection of this locus with the ovular locus outlined above will constitute a discriminatory equilibrium. Sufficient conditions for this equilibrium to exist are then that:

1. The locus defined by (6) intersects the 45 degree line at values $(\bar{q}, \bar{q})$ where $\bar{q} \in \tilde{Q}$.

2. The locus defined by (6) touches both axes on the defined $(\bar{q}_A, \bar{q}_B)$.

As the ovular locus is continuous, surrounds the 45 degree line and intersects this at (Q_0, Q_0) and (Q_1, Q_1), if the locus defined by (6) intersects the 45 degree line between (Q_0, Q_0) and (Q_1, Q_1) and it continues to the axes in both directions, then it must intersect the ovular locus. The remainder of this proof defines criteria for each of these conditions to hold.

Condition 1: The value of $\bar{q}$ that constitutes the non-discriminatory equilibrium is determined by ϕ. By assumption 2, ϕ is bounded by $(0, F(\omega))$. As (6) is continuously increasing in ϕ, if the open set $\tilde{Q}$ exists, there must be some open set $\Phi \subset (0, F(\omega))$ such that if $\phi \in \Phi$, this condition holds.

Condition 2: When $\bar{q}_A = 0$, for (6) to be satisfied, it must be that $\phi = n_B R(\bar{q}_B)$. As $R'(\bar{q}) > 0$ for all $\bar{q} \in (0, \omega)$, then it must be that $\phi \leq n_B R(\omega) = n_B F(\omega)$. Thus condition 2 requires that $\phi \leq n_g F(\omega)$ for $g \in \{A, B\}$.

Therefore, we have shown that when $\omega \in \Omega$, if $\phi \leq \min\{n_A, n_B\} F(\omega)$, then a discriminatory equilibrium exists.■

Corollary 2

The expected payoff to workers given an equilibrium cut-off application strategy is

$$EU(\bar{q}) = \frac{\omega}{2} \int_{-\infty}^{\bar{q}} \underbrace{\left(1 + \text{erf}\left(\sqrt{\frac{\eta}{2}}\left(q - \bar{q} + \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(\frac{2\bar{q}}{\omega} - 1\right)\right)\right)\right)}_{(I)} dq + \underbrace{\int_{\bar{q}}^{\infty} q f(q) dq}_{(II)} \quad (25)$$

where (I) is the probability of being hired in an equilibrium given $\underline{s}$ is self-confirming, and (II) is the income from entrepreneurship. To differentiate this with respect to the maximum productivity value of application, $\bar{q}$, I must employ the Leibniz rule to each component. Beginning with (II) first, the derivative of this with respect to $\bar{q}$ is $-\bar{q}$.

Moving to (I), when $q = \bar{q}$, the expression becomes $2\bar{q}/\omega$. Substituting $k = \text{erf}^{-1}(2\bar{q}/\omega - 1)$, $\gamma = \sqrt{2/\eta} \cdot k$, and $Z = \sqrt{\eta/2}(q - \bar{q} + \gamma)$, then we can write (I) as $1 + \text{erf}(Z)$. For the Leibniz rule, I need to compute the derivative of this with respect to $\bar{q}$ which can be written as

$$\frac{\partial}{\partial \bar{q}} \cdot (1 + \text{erf}(Z)) = \frac{\partial Z}{\partial \bar{q}} \cdot \frac{2}{\sqrt{\pi}} e^{-Z^2}$$

where $\partial Z/\partial \bar{q} = \sqrt{\eta/2}(-1 + \partial\gamma/\partial\bar{q})$. Therefore, the derivative of (I) applying the Leibniz rule can be written as

$$\frac{2\bar{q}}{\omega} + \left(\frac{d\gamma}{d\bar{q}} - 1\right)\int_{-\infty}^{\bar{q}} \frac{2}{\sqrt{\pi}}\sqrt{\frac{\eta}{2}}e^{-Z^2}dq.$$

Applying the substitution of $v = Z$, this means that $dq = \sqrt{2/\eta}\cdot dv$ and $v|_{q=\bar{q}} = k$. Therefore, the above can be rewritten as

$$\frac{2\bar{q}}{\omega} + \left(\frac{d\gamma}{d\bar{q}} - 1\right)\int_{-\infty}^{k} \frac{2}{\sqrt{\pi}}e^{-v^2}dv.$$

The integral is therefore equal to $1 + \text{erf}(k)$ which is evaluated to be to $2\bar{q}/\omega$. Moreover, we can compute $d\gamma/d\bar{q}$ as $\sqrt{2\pi/\eta}(e^{k^2}/\omega)$. Therefore, the derivative of (25) is the following

$$\frac{dEU(\bar{q})}{d\bar{q}} = \frac{\omega}{2}\left[\frac{2\bar{q}}{\omega} + \frac{2\bar{q}}{\omega}\left(\sqrt{\frac{2\pi}{\eta}}\frac{e^{k^2}}{\omega} - 1\right)\right] - \bar{q}. \tag{26}$$

This can be simplified to

$$\frac{dEU(\bar{q})}{d\bar{q}} = \bar{q}\left[\sqrt{\frac{2\pi}{\eta}}\frac{e^{k^2}}{\omega} - 1\right]. \tag{27}$$

Given Assumption 1 and that $e^{k^2} \geq 1$, the expression inside the brackets is always positive. Given that $\bar{q} \in (0, \omega)$, this derivative is always positive demonstrating that as the equilibrium cut-off productivity value increases, the aggregate earnings of any worker group increase with their equilibrium cut-off $\bar{q}_g$. ■

Proposition 3

The employer's problem is unchanged, so, given Assumptions 1 and 2 hold, the employer's optimal hiring strategy is unchanged from Proposition 1. It therefore suffices to show that pooling application strategies cannot constitute a best response when $\hat{m} < 0$.

From (1), the application strategies of workers are only pooling if they believe that the signal thresholds are equal. However, if the employer is perceived to be biased $\hat{m} < 0$, then their optimality condition is perceived to be

$$\mathbb{E}[q|\underline{s}_A, A; q \leq \bar{q}_A] = \mathbb{E}[q|\underline{s}_B, B; q \leq \bar{q}_B] - \hat{m}.$$

Since $\hat{m} < 0$, this requires the employer to set a higher expected productivity threshold for group A. By Lemma 2, $\bar{q}_A = \bar{q}_B = \bar{q}$ implies identical posterior distributions across groups. However, the optimality condition under $\hat{m} < 0$ requires $\mathbb{E}[q|\underline{s}_A, A; q \leq \bar{q}] > \mathbb{E}[q|\underline{s}_B, B; q \leq \bar{q}]$ which, by strict monotonicity of the posterior, implies that $\underline{s}_A > \underline{s}_B$. From (1), workers' cut-off application strategies depend only on the perceived thresholds, and unequal perceived thresholds generate unequal cut-offs by the uniqueness result in Lemma 1. Thus workers cannot perceive pooling strategies as a best response when $\hat{m} < 0$, so the non-discriminatory equilibrium is not a Bayesian Nash Equilibrium under Definition 2. ■

Proposition 4

Under $\hat{\psi} = 0$ and $\hat{m} < 0$, workers believe the employer sets hiring thresholds satisfying $\hat{m} = \underline{s}_B - \underline{s}_A$. Substituting the threshold function $\underline{s}(\bar{q})$ derived above gives the perceived optimality locus

$$\hat{m} = \left(\bar{q}_B - \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(\frac{2\bar{q}_B}{\omega} - 1\right)\right) - \left(\bar{q}_A - \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(\frac{2\bar{q}_A}{\omega} - 1\right)\right). \tag{28}$$

Since $\underline{s}(\bar{q})$ is strictly decreasing in $\bar{q}$, the condition that $\hat{m} < 0$ requires $\underline{s}_A > \underline{s}_B$, which in turn requires $\bar{q}_A < \bar{q}_B$. Thus, the locus defined by (28) is strictly above the 45 degree line on the $(\bar{q}_A, \bar{q}_B)$ plane.

Fixing $\bar{q}_B$, a lower $\hat{m}$ requires a higher $\underline{s}_A$ which, by strict monotonicity of $\underline{s}$, requires a lower $\bar{q}_A$. Thus, for any fixed $\bar{q}_B$, the corresponding $\bar{q}_A$ is strictly increasing in $\hat{m}$.

By the definition of the satiation condition, (13), as $\bar{q}_A$ decreases, $\bar{q}_B$ must increase to satisfy the condition. Therefore, if the non-discriminatory equilibrium exists, then the satiation locus (13) is decreasing on the $(\bar{q}_A, \bar{q}_B)$ plane and crosses the 45 degree line at the non-discriminatory equilibrium $\bar{q}^*$. For $\hat{m} = 0$, the perceived optimality locus (28) is the 45 degree line so the two loci intersect at $\bar{q}^*$. Given the continuity of (28) in $\hat{m}$, (28) approaches the 45-degree line asymptotically as $\hat{m} \to 0$.

As $\underline{s}(\bar{q}_A) \to +\infty$, $\bar{q}_A \to 0$, so the condition $\underline{s}(\bar{q}_B) = \hat{m} + \underline{s}(\bar{q}_A)$ requires $\underline{s}(\bar{q}_B) \to +\infty$, hence $\bar{q}_B \to 0$. (28) therefore approaches the origin in the limit.

At the upper limit, as $\underline{s}(\bar{q}_A) \to -\infty$, $\bar{q}_A \to \omega$, so the condition $\underline{s}(\bar{q}_B) = \hat{m} + \underline{s}(\bar{q}_A)$ requires $\underline{s}(\bar{q}_B) \to -\infty$ which requires $\bar{q}_B \to \omega$. Thus, (28) approaches (ω, ω) in the upper limit. However, given that $\bar{q} \in (0, \omega)$, the locus cannot intersect either of these limiting points.

Given the above, the locus prescribed by (28) lies between $(0, 0)$ and (ω, ω), touching neither point, above the 45-degree line when $\hat{m} < 0$. Moreover, at intermediate values of $\bar{q}_B \in (0, \omega)$, the prescribed value of $\bar{q}_A$ is increases with $\hat{m}$. In the limit of $\hat{m} \to -\infty$, for any prescribed $\bar{q}_B$, it must be that $\underline{s}(\bar{q}_A) = \infty$ which implies that for any $\bar{q}_B$ below ω, the corresponding value of $\bar{q}_A$ must be 0. As $\underline{s}(\bar{q})$ is a continuous function in $\bar{q}$ and (28) is continuous in $\hat{m}$, the intersection point of the satiation locus, (13), and the perceived optimality constraint, (28), moves further from the 45-degree line as $\hat{m}$ decreases until it intersects at some point $(0, \bar{q}_B^{\dagger})$ when $\hat{m} = -\infty$. Therefore, as the non-discriminatory equilibrium defined by Proposition 2 lies along the satiation locus, (13), between the 45-degree line and the bound $(0, \bar{q}_B^{\dagger})$, by the intermediate value theorem, there must be some value of $\hat{m} < 0$ such that (28) and (13) intersect at the non-discriminatory equilibrium defined by Proposition 2.■

C On the cause of misattribution

3a Theory

Section 2 establishes that misattribution sustains discrimination. I now consider the potential causes of misattribution, identifying two distinct sources. Together, Propositions 1 and 3 demonstrate that the non-discriminatory equilibrium can only be sustained when the employer is perceived as unbiased and all workers are believed to play pooling application strategies. Proposition 4 implies that workers may attribute the discriminatory equilibrium to the employer's bias or to separating application strategies and the employer's sophistication. If workers correctly perceive the strategies of their colleagues but believe that the employer does not respond to separating application strategies then they may misattribute the source of discrimination. This is the first explanation outlined above.

The second source of misattribution does not require workers to misperceive employer sophistication, but instead to underestimate how much other workers condition their strategies on anticipated discrimination. Suppose that workers hold correct beliefs about the level of sophistication of the employer, but underestimate the anticipatory behaviour of other workers, $|\bar{q}_A - \bar{q}_B| > |\hat{\bar{q}}_A - \hat{\bar{q}}_B|$. In other words, the workers are *cursed* in that they underestimate the distance between the application strategies of the two groups.

Taking an extreme case of cursedness where workers incorrectly believe that no other workers condition their application strategies on potential discrimination, $\hat{\bar{q}}_A = \hat{\bar{q}}_B$, the only equilibrium attainable is the discriminatory equilibrium outlined above.

Proposition 5: Cursed Equilibrium. Suppose that Assumption 1 holds and that workers' beliefs of employer type are $\hat{\psi} = 1$. Additionally suppose that workers incorrectly believe that $\bar{q}_A = \bar{q}_B$, then there exists some $\hat{m} < 0$ such that the equilibrium conditions in Definition 2 hold.
Proof. Under $\hat{\psi} = 1$ and the cursedness assumption $\hat{\bar{q}}_A = \hat{\bar{q}}_B = \hat{\bar{q}}$, workers believe the employer is sophisticated but perceive both groups as playing the same cut-off strategy. By Lemma 2, a sophisticated employer facing equal perceived application strategies sets equal signal thresholds absent any bias, i.e. $\hat{\underline{s}}_A = \hat{\underline{s}}_B$ when $\hat{m} = 0$. Under perceived bias $\hat{m} < 0$, the sophisticated employer's optimality condition requires

$$\mathbb{E}[q|\underline{s}_A, A; q \leq \bar{q}_A] = \mathbb{E}[q|\underline{s}_B, B; q \leq \bar{q}_B] - \hat{m}.$$

Since workers believe that $\hat{q}_A = \hat{q}_B$, the posterior distributions are perceived identical across groups. By the strict monotonicity of the posterior in the signal, this requires $\hat{\underline{s}}_A = \hat{\underline{s}}_B - \hat{m}$, which is exactly the perceived relationship under $\hat{\psi} = 0$ in Proposition 4. The remainder of the proof: that the satiation and perceived optimality loci intersect at the discriminatory equilibrium from Proposition 2 for some $\hat{m} < 0$, then follows by the argument of Proposition 4.■

One can interpret the cursed equilibrium in the following way. Suppose that workers observe the employer only hired B workers. Workers may then anticipate discrimination against A workers meaning that separating application strategies may be optimal strategies. Consequently, the statistical discriminator employer hires only B workers again. If workers incorrectly believe that all other workers played pooling strategies, then the only belief of employer type through which they can rationalise this outcome is that the employer is a taste-based discriminator. However, if they correctly perceive the strategies of other workers, then they may foresee that a statistical discriminator responding to such applications may produce the same hiring outcomes.

3b Empirical evidence

Applying the theoretical intuition to my experiment, the first proposed mechanism causing misattribution would be consistent with individuals being aware of the true strategies of others, but struggling with the conceptual difficulty in correcting for such selection. In other words, workers may not accurately predict the Bayesian response to separating application strategies. The mechanism proposed in Proposition 5 proposes that individuals have an incorrect view of the data generating process:they are unaware that the reason they do not observe non-discrimination is because other workers play the same strategy as them.

In this section, I present evidence from the laboratory experiment that misattribution is not driven by cursedness. To do, I draw evidence from both the baseline beliefs of others' application strategies and data from the *strategic* treatment. In this treatment, participants were informed of the true maximum productivity values of green and purple applicants after they stated their belief of such in decision 2 and before they reported the probability that the employer was type A in decision 3. If misattribution is driven by cursedness, the underestimation of the degree to which others condition their strategies on anticipated discrimination, then revealing the true application cut-offs should reduce misattribution even when hiring outcomes remain unchanged.

Baseline data In each round of the experiment, participants were asked to state their beliefs of the maximum productivity value of green and purple applicants in their society. In the pooled baseline data which includes the baseline treatment and the first four rounds of the *revealed* treatment, the

Table C.1: Average guessed maximum productivity of applicants

	(1)	(2)	(3)	(4)
Subjective max. productivity of				
Green applicants	70.486	70.828	67.537	71.481
	(1.365)	(1.509)	(1.691)	(2.582)
Purple applicants	77.719	78.266	77.560	83.462
	(1.331)	(1.471)	(1.645)	(2.429)
Two-sided t-test				
Difference	-7.232***	-7.434***	-10.024***	-11.982***
	(1.119)	(1.281)	(1.487)	(2.375)
Share predicting pooling strategy	0.236	0.224	0.205	0.172
Excludes round 1		✓	✓	✓
Average purple hires ≥ 50%			✓	✓
Believe type A > 50%				✓
Observations	1536	1280	976	493

Notes: Average guesses of the maximum productivity value of green and purple applicants in one's society. Full sample refers to all participants in the baseline and first four rounds of the *revealed* treatment. Column (2) excludes guesses made in round 1. Column (3) excludes participants who observe a hiring history from the previous round in which less than 50% of hires going to purple candidates. Column (4) includes only those whose previous round's belief of the probability that the employer is type A was above 50%, the Bayesian benchmark. Standard errors are displayed in parentheses. Share predicting pooling strategies maps the share of participant-round observations in which the reported maximum productivity of green and purple applicants are equivalent. Stars of the difference correspond to the significance of two-sided t-tests to test the equivalence of guesses about green and purple applicants: *p<0.10, **p<0.05, ***p<0.01.

average belief of green applicants' maximum productivity was 70.5 and that of purple applicants was 77.7. These are significantly above the true values of 57.6 and 66.1 respectively (two-sided t-tests $p < 0.001$).[32]

The majority of participants believe that application strategies are colour-dependent. Table C.1 demonstrates that the difference between the average reported maximum productivity value of green and purple applicants is statistically significant and this difference remains when excluding round 1 elicitations and those who observe hiring histories that favour green applicants. Importantly, the difference between average guesses increases when focusing only on participants who misattribute the source of discrimination: those who report a subjective probability of the employer being type A above 50%.[33] Moreover, a pooling strategy is predicted in only 23.6% of all participant-round observations and in only 18% in the observations which encountered discrimination. This is evidence against cursedness being the cause of misattribution in the baseline.

Table C.2 presents results from two sets of regressions which formally show that the subjective probability that the employer is type A is positively correlated with the perceived likelihood of separating application strategies. All regressions control for the hiring history and standard errors are clustered at the society level. In columns 1 and 2, the presented estimates correspond to linear

[32]To compute these t-test, I exclude participants in society-rounds for which no participants of a colour applied.

[33]Table E.12 demonstrates that beliefs also do not converge over each round as a difference remains in each round which is statistically significant in all but rounds 4 and 5.

Table C.2: Effect of believed probability that the employer is type A on cursedness

Dependent variable	Difference between purple and green guesses (OLS)		Green and purple guesses are equivalent (AME)	
	(1)	(2)	(3)	(4)
Subj. prob. of type A	18.452***	9.937**	-0.209***	-0.210***
(Previous round)	(5.067)	(3.592)	(0.049)	(0.060)
Share of hires purple		34.213***		0.007
(Previous round)		(7.089)		(0.082)
Session FEs	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓
Round FEs	✓	✓	✓	✓
Observations	1280	1280	1254	1254
R^2	0.034	0.045		

Notes: Columns (1) and (2) estimate a linear model on the effect of the previous period's subjective probability that the employer is type A on the difference between the guessed maximum productivity value of purple and green applicants. Columns (3) and (4) show average marginal effects from logit models on the effect of the previous period's subjective probability that the employer is type A on the propensity of the participant to state equivalent guesses of the maximum productivity value of green and purple applicants. The sample is restricted to individuals in the baseline set of data from round 2 onwards. All models include session fixed effects, round fixed effects and individual controls for age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01.

models which test whether last period's subjective probability of the employer being type A affects the difference between one's guesses of the maximum productivity of purple and green hires. The positive coefficient estimates indicate that the more likely a participant states the employer is type A, the more separated they perceive application strategies to be.

Columns 3 and 4 of Table C.2 present average marginal effects from a logistic regression of the indicator that a subject perceives application strategies to be pooled on their subjective probability of the employer being type A. The estimated coefficients are negative and statistically significant, demonstrating that the more likely a participant reports the employer to be type A, the taste-based discriminator, the less likely the participant is to report that the maximum productivities of green and purple applicants are equal. The coefficient point estimate intimates that at a given hiring history, a 10 percentage point increase in the subjective probability that the employer is type A decreases the probability that the participant reports a pooling strategy by 2.1 percentage points.

To summarise, individuals who perceive maximum productivities of applicants from each group to be closer together are less likely to believe that the employer is a taste-based discriminator. However, there is a reverse-causality concern in interpreting this relationship as causal: it may be that participants perceive others are more likely to play the unique optimal separating application strategies when the employer is perceived to be type A. To attest to this concern, I analyse the effect of the *strategic* treatment on reported probabilities of employer type.

The causal effect of strategic uncertainty on misattribution In eight randomly selected societies, I implemented a treatment arm *strategy* which proceeded identically to the baseline except that participants were informed of the true maximum productivity values of green and purple applicants in their society when they elicited their belief of employer type. This additional information effectively reduced the strategic uncertainty of the game and thus the potential for participants to be cursed. [34]

As intended, the *strategic* treatment substantially corrected beliefs of strategies compared to the baseline. After receiving their first information signal, participants in the *strategic* treatment believed that, on average, the maximum productivity value of green applicants was 51.6 and 54.9 for purple applicants as presented in Table E.13. On average, the reported difference between the maximum productivity value of purple and green applicants was less in the *strategic* treatment than in the pooled baseline data (3.390 vs. 7.438 two-sided t-test $t = 1.980$ $p = 0.048$). However, the share of participants reporting equivalent maximum productivity values of purple and green applicants is lower in the *strategic* treatment than in the pooled baseline.[35] Therefore, while participants in the *strategic* treatment report a lower difference between maximum productivities of green and purple applicants than participants in the baseline data, they are less likely to report that these maximum productivities are equal.

To formally test whether reducing the scope for misperception of others' strategies by being informed of the true difference between maximum productivity values of applicants from each group reduces the subjective probability that an employer is type A, I estimate a series of linear OLS models to isolate the treatment effect of the *strategic* condition. To control for non-linearities in the treatment-relevant information provision, I include only participants in rounds for which at least one green and at least one purple application occurs in their society. Participant observations excluded from this sample are those in rounds for which the no green and/or purple participants from their society applied, and thus, if they were in the *strategic* treatment, would have not received any quantifiable information of others' strategies. The results presented in Table C.3 provide evidence against the conjecture that reducing strategic uncertainty reduces the scope for misattribution: the coefficient estimates of the *strategic* dummy variable are not negative. In fact, when excluding participants facing hiring histories in which green workers are favoured, the coefficient estimate is positive and significant at the 1% level (column 5).

In specification 6 of Table C.3, I only include participant observations for which the reported difference in maximum productivity values of purple and green applicants is less than the true difference. Those in this sample are participants who would have been informed that the true difference between application strategies is larger than perceived if they were in the *strategic* treatment. Interpreting the *strategic* treatment coefficient in this specification, being informed that the maximum productivities of purple and green applicants in that society and round increases the subjective probability of the employer being type A by 8.6 percentage points. This result is evidence that when the hiring history does not favour green workers, a reduction of strategic uncertainty does not reduce the scope of misattribution, it increases it.

To control for the information value in the *strategic* treatment, I regress the subjective probability of the employer being type A on the perceived and true difference between the maximum productivity values of purple and green applicants in Table E.14. To analyse the effect of reveal-

[34]In some cases (N=48), no purple or green participants applied in a society. If this was the case, participants were informed.

[35]Among participants who observe a hiring history of at least 50% of hires going to purple workers and who report a subjective probability that the employer is type A of more than 50% (who overestimate the likelihood that the employer is a taste-based discriminator compared to the Bayesian benchmark), the share reporting equal maximum productivity values of green and purple applicants from round 2 onwards is 17.2% in the pooled baseline data and 10.2% in the *strategic* treatment (two-sided t-test $t = 2.577$ $p = 0.01$).

Table C.3: Effect of *strategic* treatment on misattribution

Dependent variable	Subjective probability that employer is type A					
	(1)	(2)	(3)	(4)	(5)	(6)
Share of hires purple	0.606***	0.698***	0.722***	0.721***	1.008***	1.062***
	(0.065)	(0.057)	(0.052)	(0.050)	(0.096)	(0.144)
Strategic	-0.008	0.016	0.021	0.026	0.059***	0.086**
	(0.030)	(0.020)	(0.020)	(0.019)	(0.020)	(0.035)
Session FEs		✓	✓	✓	✓	✓
Individual controls			✓	✓	✓	✓
Round FEs				✓	✓	✓
Excl. share of purple hires <50%					✓	✓
True distance>perceived distance						✓
Baseline mean	0.455	0.455	0.455	0.455	0.505	0.524
Observations	2512	2512	2512	2512	1824	902
R^2	0.166	0.207	0.252	0.252	0.254	0.250

Notes: Linear fixed effects models estimating the treatment effect of the *strategic* arm relative to the baseline on the subjective probability that the employer is type A controlling for current hiring histories and guesses of maximum productivity values of green and purple applicants. The sample excludes participants in the *revealed* treatment after round 4 and all participants in rounds for which there is a colour in their society in which no worker applies. Specifications 4, 5 and 6 include individual controls for age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Specifications 5 and 6 excludes participants facing a hiring history in which the share of purple hires is less than 50%. Specification 6 incudes only individuals who believe the difference between the maximum productivity of purple and green applicants is less than the true difference. Baseline mean is the average subjective probability of the employer being type A where the average share of hires going to purple candidates is 57.9% (columns 1-4), 66.8% (column 5) and 69.1% (column 6). Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01.

ing this information to participants on their beliefs, I interact the true difference with the *strategic* treatment dummy variable and exclude participant observations for which green or purple workers were absent from their society's applicant pool in the specific round. Across all specifications, I do not find evidence that the information affects the subjective probability of the employer being type A. Importantly, this result holds when I include only participants who underestimate the difference between the maximum productivity of purple and green applicants. Thus, when individuals are informed that application strategies are more separated than they perceive them to be, their subjective probability of the employer being type A remains unchanged.

Result 5: The misinterpretation of application strategies does not cause the misattribution of discrimination source.

D Practical procedure of the employer's updating

In practise, employer type B conducted two sets of Bayesian inference. In each round, it first predicted the underlying productivity values of applicants given their colour, productivity value and its belief of participant strategies: the maximum productivity value at which participants from each colour would apply. Second, it would then update its belief of strategies using the distribution of signals observed from each colour.

Type B assumes that a participants from group g adopt a cutoff application strategy and only apply if their productivity value is below some productivity value threshold $\bar{q}_g$. The Bayesian type predicts the underlying productivity value using its prior belief of $\bar{q}_g$, the observed signal, the known concentration τ of productivity and signal precision η just as in the theoretical model.

After observing a full round's signals for a given group, type B updates its belief about that colour's application threshold $\bar{q}_g$. This is treated as a Bayesian updating problem over a discretised grid of candidate threshold values. Employer B holds a prior belief $\bar{q}_g \sim N(m_g, v_g^2)$, calibrated separately for each group and carried forward from the previous round's posterior.

For each candidate value of $\bar{q}_g$ on the grid, the model computed the implied truncated-normal moments $(\mu_{trunc}(\bar{q}_g), \tau_{trunc}(\bar{q}_g))$ where $\mu_{trunc}(\bar{q}_g)$ is the mean value of q given a truncation at candidate value $\bar{q}_g$ and $\tau_{trunc}(\bar{q}_g)$ is the corresponding concentration. It then evaluated the likelihood of that round's observed signals under a normal distribution with mean $\mu_{trunc}(\bar{q}_g)$ and concentration $1/\tau_{trunc}(\bar{q}_g) + 1/\eta$. Critically, this likelihood is a joint likelihood across all signals observed from that group in the round, not for each signal individually,

$$Log\mathcal{L}(\bar{q}_g) = \sum_{i=1}^{K} log\gamma(s_i|\mu_{trunc}(\bar{q}_G), 1/\tau(\bar{q}_g) + 1/\eta) \tag{29}$$

where K is the number of applicants from group g in the round. Combining this with the log-prior carried forward and normalising numerically using a trapezoidal integration over the grid, yields the full posterior distribution over $\bar{q}_g$ from which type B extracts a posterior mean and standard deviation, (m_g, v_g). This posterior mean becomes the prior mean for the group entering the next round.

To calibrate updating, I simulated experimental data to ensure that the employer's belief of strategies could converge promptly to either equilibria should participants converge their strategies to such while permitting some lag to allow participant learning. Thus, I set the standard deviation of this initial belief to be 15. Figures D.1 and D.2 show simulations of this procedure when the economy remains in a discriminatory and non-discriminatory equilibrium respectively. These simulations were conducted prior to the first session in which it was determined that for financial reasons, only 8 rounds could be conducted instead of the intended 12. Due to time constraints with the laboratory, this amendment had to be made overnight and so further simulations were not possible.

Robustness of the procedure to shocks

A shock that may occur in the updating process concerns the introduction of a quota. Figures D.3 and D.4 present simulations of the updating procedure after a successful and unsuccessful quota respectively. A successful quota refers to one in which participants successfully converge to non-discriminatory application strategies following the quota's removal. As evidenced, the updating procedure is sensitive to such a coordinated shift in strategies.

A natural concern with the updating procedure is whether a single participant's atypical behaviour, for instance, one participant applying with an unusually high productivity value contrary

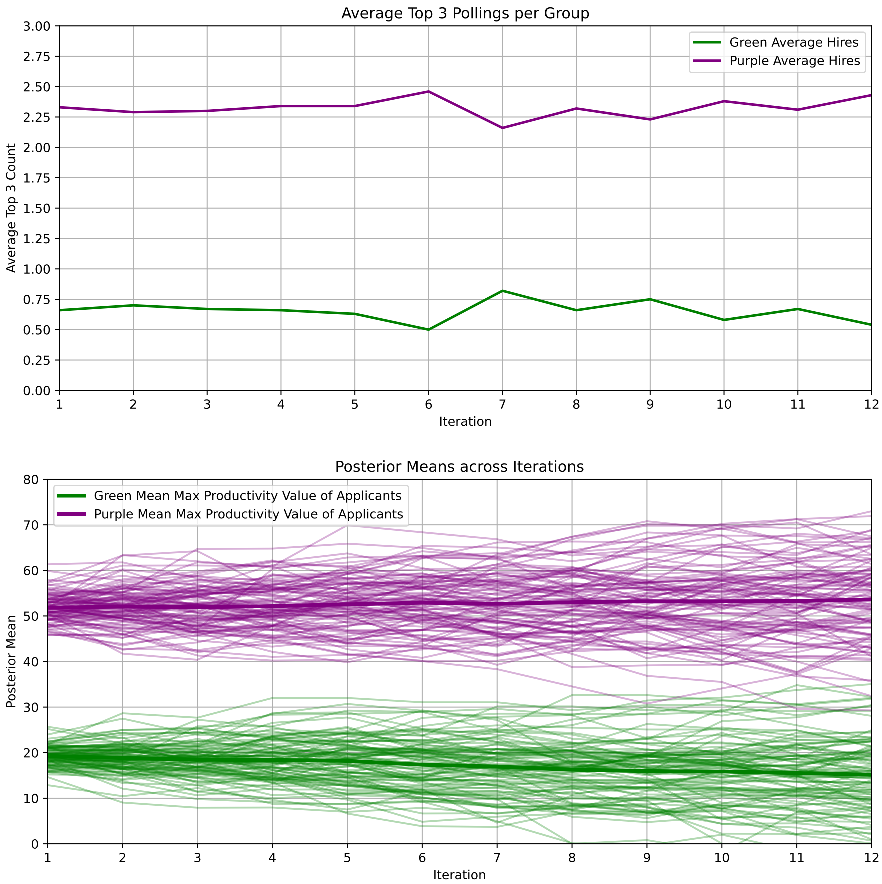


Figure D.1: Discriminatory equilibrium

Note: Simulated experiment outcomes with starting point outlined in Figure 3 and prior standard deviation 15. Simulated application strategy is congruent with DE1 where green workers apply if and only if their productivity value is less than or equal to 1 and purple workers apply if and only if their productivity value is less than or equal to 58. Upper graph shows the simulated average rate of hiring green and purple workers. Lower graph shows the simulations of posterior means of participant application strategies by colour. Thick lines are the average over simulations.

to the rational strategy, could disproportionately distort type B's belief about the entire group. Three features of the procedure jointly mitigate this risk.

First, since the likelihood function in equation (29) aggregates log-densities across all signals observed from a group in a given round, one atypical signal is evaluated jointly alongside the remaining signals from co-applicants, rather than driving the update in isolation. Therefore, the influence of the atypical signal is diluted by however many other applicants from that group appear in the same round.

Second, the prior belief $N(m_g, v_g^2)$ carried over from the previous round acts as an anchor. Since the posterior is proportional to the product of the prior and the likelihood, a moderately informative prior exerts continued pull toward the previous round's belief. This means that evidence from a single round, however extreme, cannot move the posterior arbitrarily far.

Third, the signal-generating process itself embeds substantial idiosyncratic noise: with my parametrisation, the signal noise precision implies a standard deviation of 81.85. Since the noise term enters the likelihood's variance directly, individual signals are inherently unreliable evidence about the underlying threshold. This also limits the incentive for atypical behaviour in an effort to

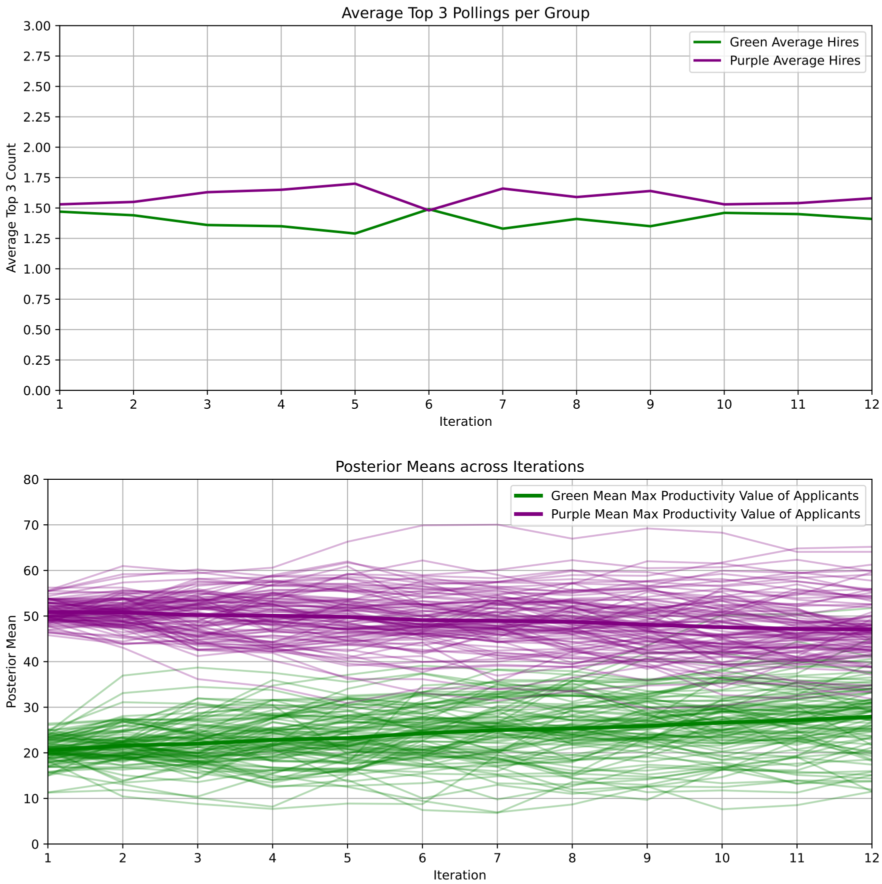


Figure D.2: Non-discriminatory equilibrium

Note: Simulated experiment outcomes with starting point outlined in Figure 3 and prior standard deviation 15. Simulated application strategy is congruent with NE where both groups of workers apply if and only if their productivity value is less than or equal to 38. Upper graph shows the simulated average rate of hiring green and purple workers. Lower graph shows the simulations of posterior means of participant application strategies by colour. Thick lines are the average over simulations.

nudge the belief updating process: a participant cannot guarantee the production of a high signal and cannot guarantee that next round they draw a productivity value sufficient to repeat the atypical behaviour.

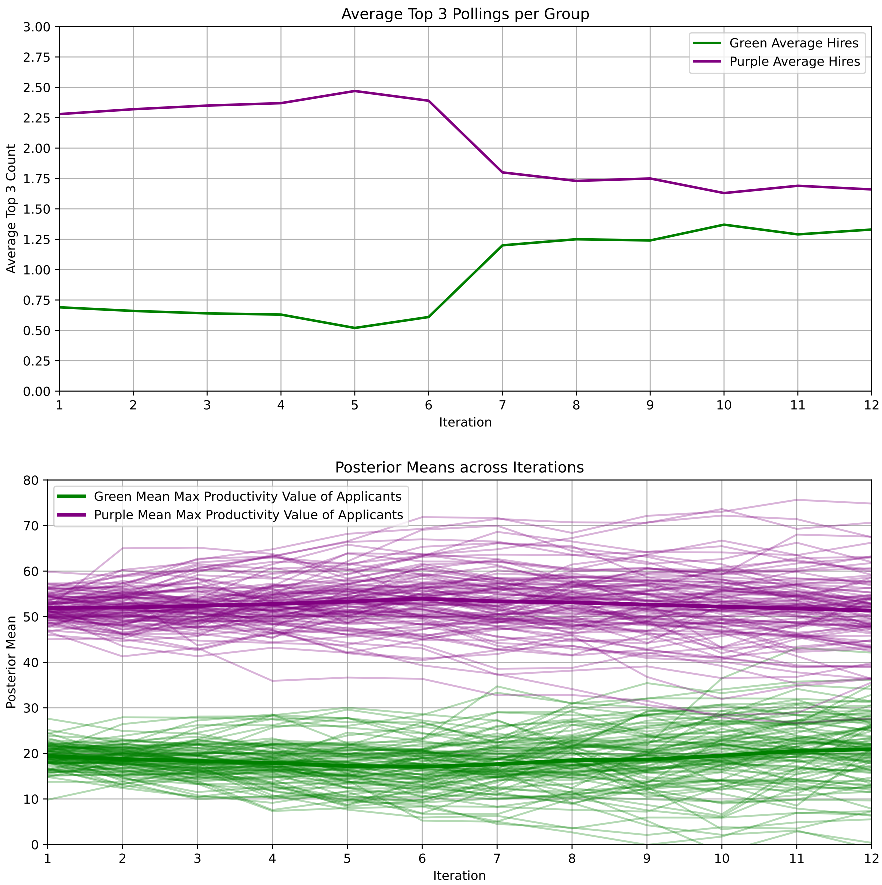


Figure D.3: Successful Quota in round 7

Note: Simulated experiment outcomes with starting point outlined in Figure 3 and prior standard deviation 15. Simulated application strategy is congruent with DE1 in the first 6 rounds and NE in the last 6 rounds. This is akin to a quota being enforced in round 7 and thereafter strategies converge to the non-discrimination equilibrium. Upper graph shows the simulated average rate of hiring green and purple workers. Lower graph shows the simulations of posterior means of participant application strategies by colour. Thick lines are the average over simulations.

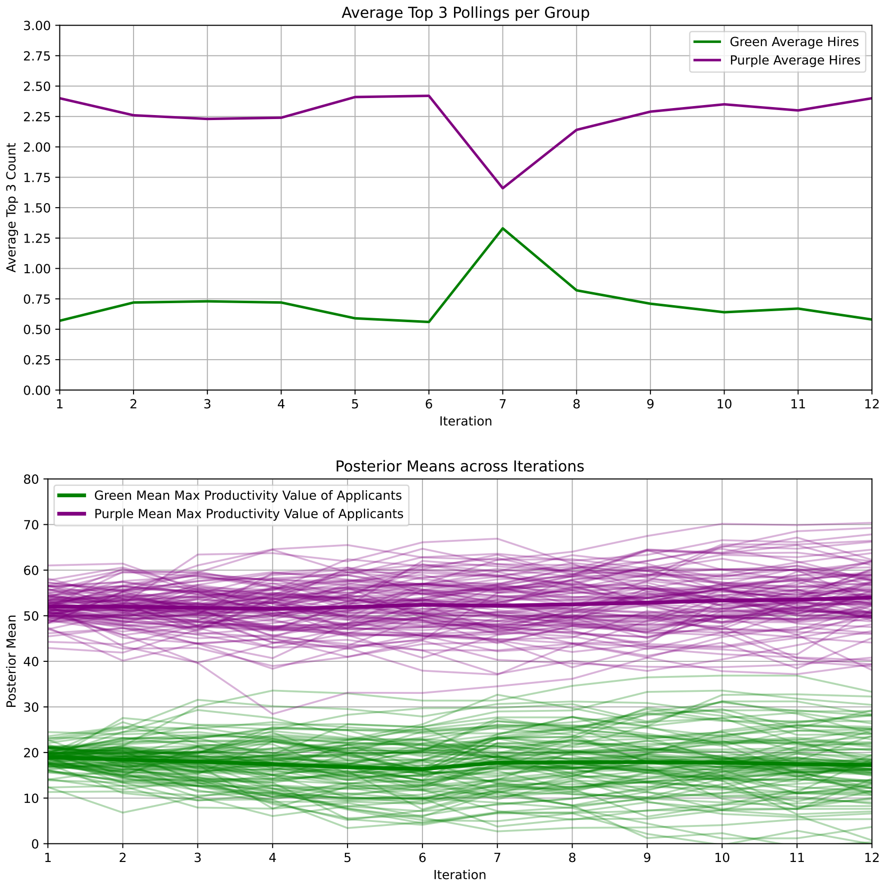


Figure D.4: Unsuccessful Quota in round 7

Note: Simulated experiment outcomes with starting point outlined in Figure 3 and prior standard deviation 15. Simulated application strategy is congruent with DE1 except in round 7 where a quota is implemented and NE is enforced. Upper graph shows the simulated average rate of hiring green and purple workers. Lower graph shows the simulations of posterior means of participant application strategies by colour. Thick lines are the average over simulations.

E Additional results referenced in text

Figures

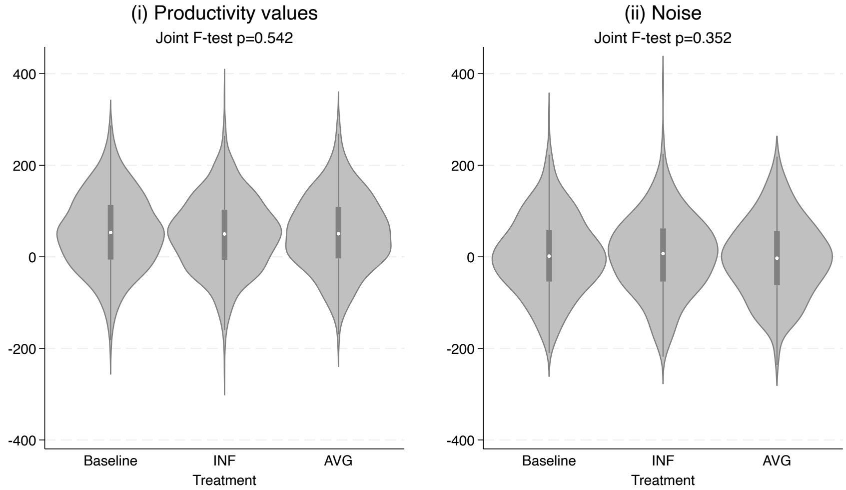


Figure E.1: Distribution of random draws across treatments

Note: This graph shows the distribution of (i) productivity values and (ii) noise across treatments. All rounds are pooled together. P-values are computed by joint F-test of mean difference.

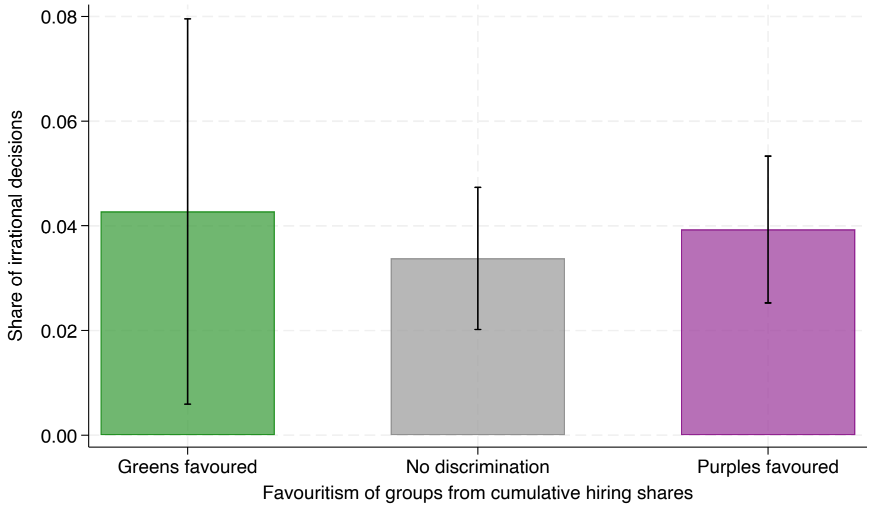


Figure E.2: Share of decisions that are irrational according to the favouritisim implied by hiring decisions

Note: Irrational decisions refer to instances in which a participant draws a productivity value of 90 or more and applies for a job or one of less than 0 and does not apply. "Greens favoured" refers to the occasions where more than 2/3 of the past hiring decisions have been to green workers. "No discrim" refers to occasions where this cumulative average shows that the share of past hiring decisions in favour of purple workers has been between 1/3 and 2/3. "Purples favoured" refers to the occasions where 2/3 or more of the past hiring decisions have been to purple workers. 95% confidence intervals presented.

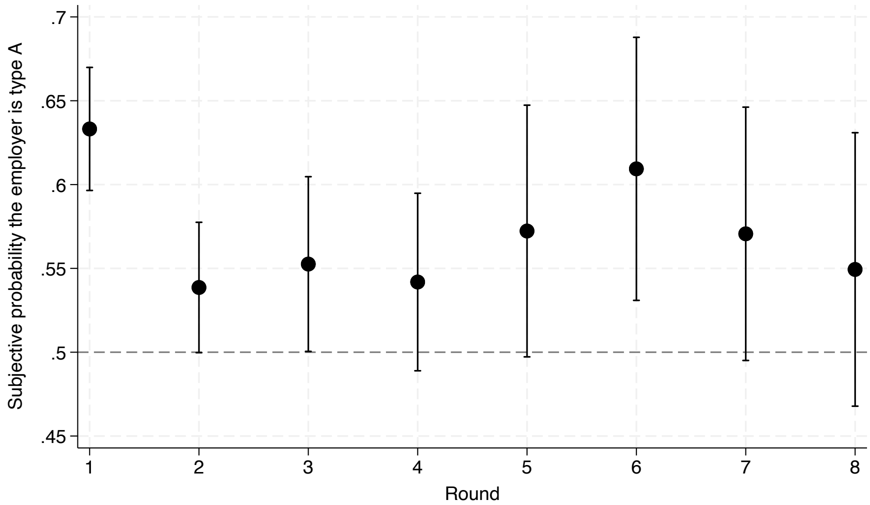


Figure E.3: Average subjective probability that the employer is type A by round when purple workers are favoured in the hiring history

Note: This graph shows the average believed probability that the employer is type A (the taste-based discriminator) is in use by individuals observing a hiring history in which at least 2/3 of the hires are purple. Error bars show 95% confidence intervals.

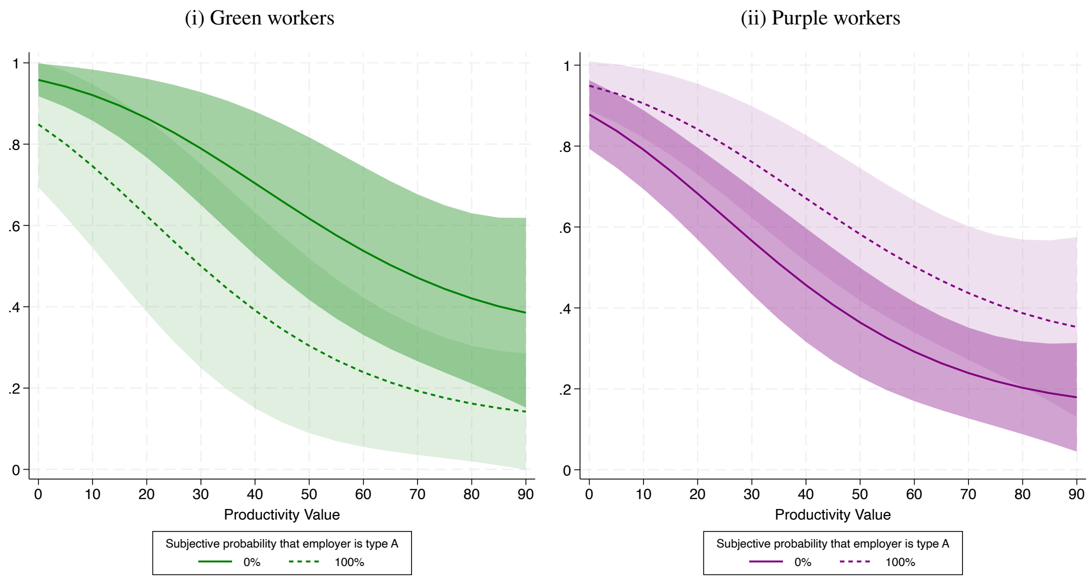


Figure E.4: Predicted probability of application at alternative attributions

Note: Each curve plots the average adjusted predicted probability of application, $\hat{P}(\text{apply} = 1 \mid \text{prod} = q, \text{belief} = b)$, for each colour group of workers, computed from a logit model estimated using participants in rounds they draw a productivity value in the set $[0, 90)$. Predicted probabilities are evaluated over a grid of productivity values in this range (step = 5) at two fixed values of the participant's subjective belief that the employer is type A $\{0, 1\}$. The logit model includes controls for previous round's subjective probability of the employer being type A, a linear and quadratic term for productivity value, session and round fixed effects as well as individual controls including age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelors), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997) investment game. Shaded bands denote 95% confidence intervals based on standard errors clustered at the society level.

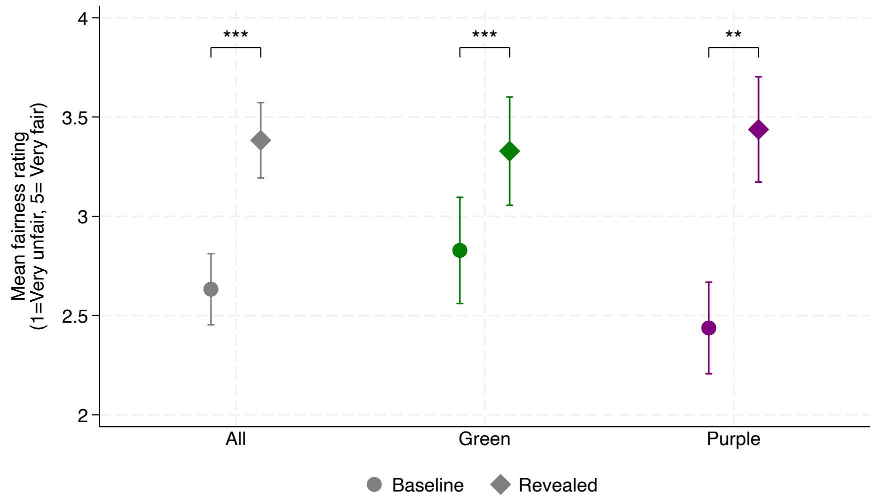


Figure E.5: Fairness perception of hiring process after round 4

Note: This graph shows the average response to the question *To what extent do you consider the hiring procedure has been fair thus far?* asked after round 8. Responses are coded to a 5-point Likert scale where 1 is *very unfair* and 5 is *very fair*. Individuals are asked this question after being informed of the true employer type, if in *revealed* treatment. Error bars show 95% confidence intervals.

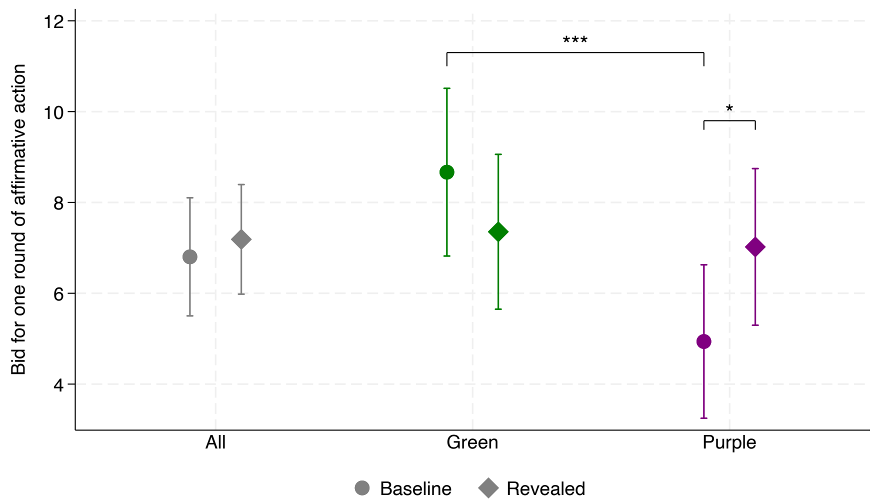


Figure E.6: Demand for temporary affirmative action conditional on hiring history

Note: This graph shows the average bid for one period of affirmative policy to be enacted in round 5 conditional on the hiring history observed at round 4 showing at least 50% of previous hires going to purple workers. Participants could bid between 0 and 20 tokens and place their bid after being informed of the true employer type, if in *revealed* treatment. Error bars show 95% confidence intervals. Stars show the significance of the difference computed from a two-sided t-test: *p<0.10, **p<0.05, ***p<0.01.

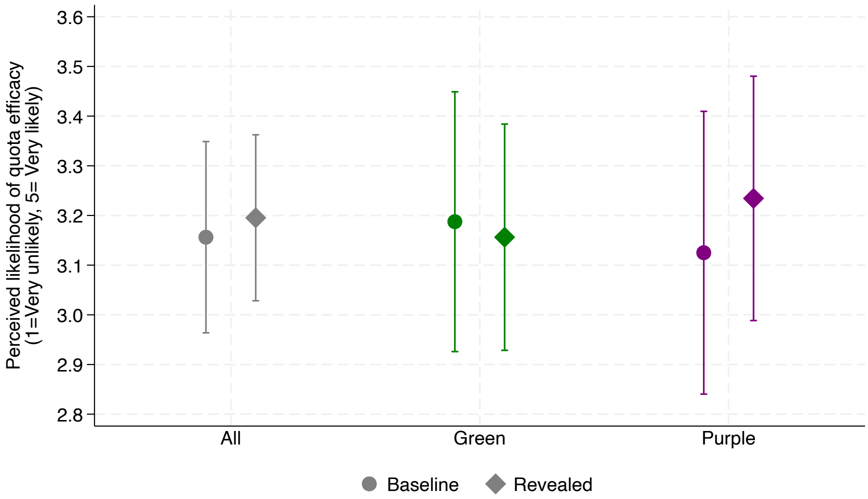


Figure E.7: Ex-ante perceived likelihood of temporary affirmative action efficacy

Note: This graph shows the perceived likelihood of the quota being effective in increasing the share of green hires as elicited in the midline survey. Responses are recorded using a 5-point Likert scale in which 1 is equal to very unlikely and 5 is equal to very likely. Error bars show 95% confidence intervals.

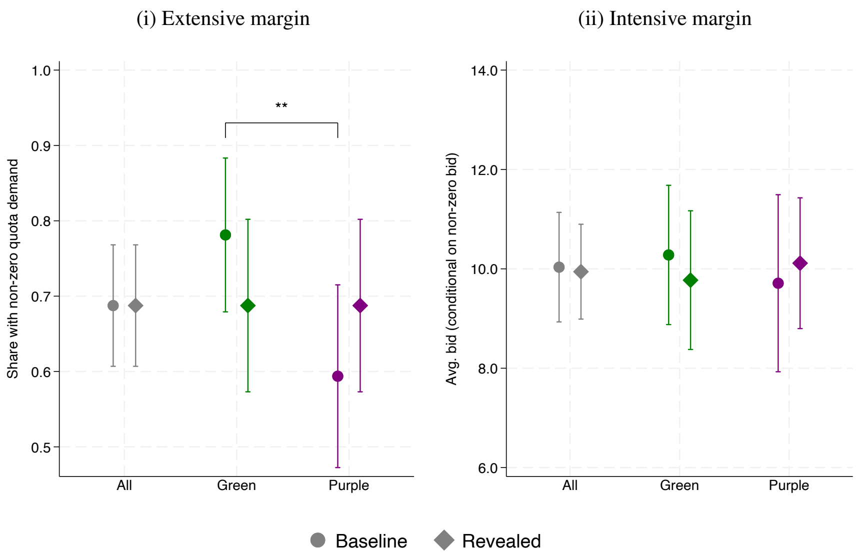


Figure E.8: Extensive and Intensive margins of demand for a quota

Note: These graphs show the demand for one round of a quota as described by the bid on the (i) extensive and (ii) intensive margins. Panel (i) shows the proportion of participants with a non-zero bid for the quota and panel (ii) shows the average bid for the quota conditional on the bid not being 0. Error bars show 95% confidence intervals. Stars show the significance of the difference computed from a two-sided t-test: *p<0.10, **p<0.05, ***p<0.01.

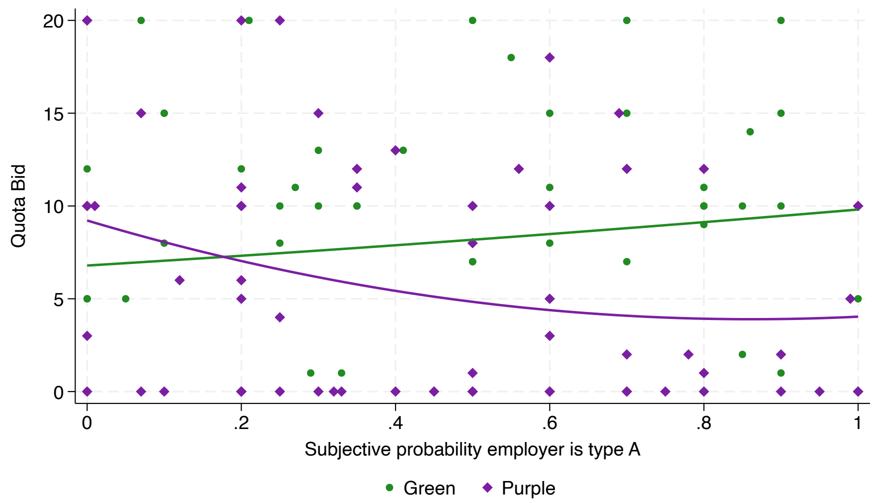


Figure E.9: Bids for temporary affirmative action by subjective probability of employer being type A

Note: This graph plots the bids for one period of a quota by the subjective probability of the employer being type A by colour for participants in the baseline treatment. A quadratic best fit line is imposed on the scatter plot.

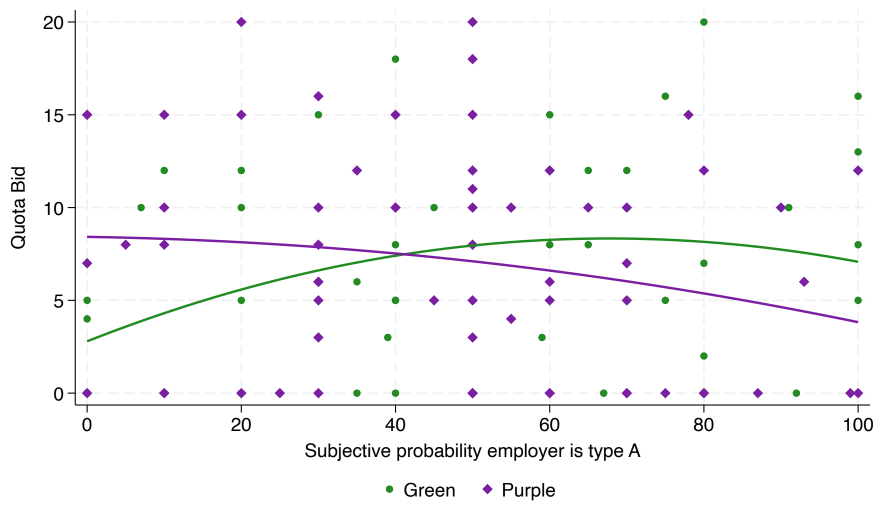


Figure E.10: Bids for temporary affirmative action by subjective probability of employer being type A (*revealed* treatment)

Note: This graph plots the bids for one period of a quota by the subjective probability of the employer being type A by colour for participants in the *revealed* treatment. A quadratic best fit line is imposed on the scatter plot.

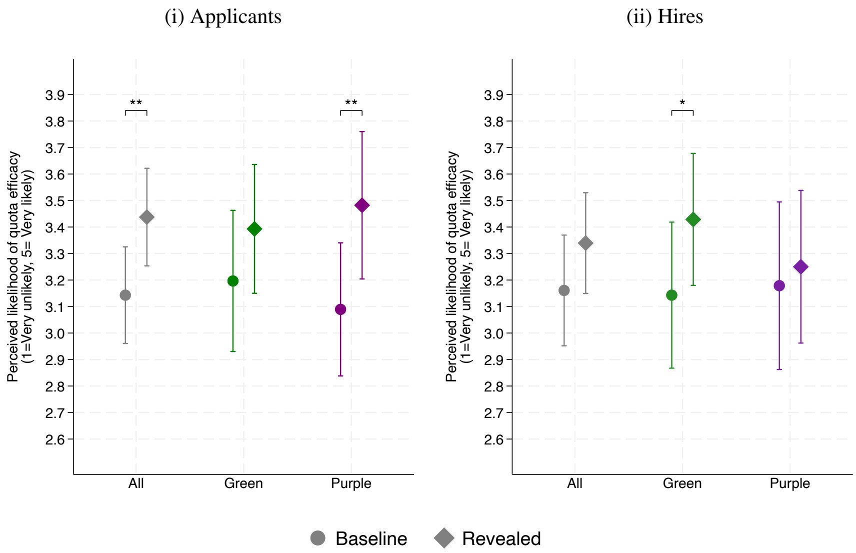


Figure E.11: Ex-post perceived likelihood of temporary affirmative action efficacy, participants who did not experience a quota

Note: This graph shows the perceived likelihood of the quota being effective in increasing the share of green (i) hires and (ii) applicants as elicited in the final survey. Participants who experienced a quota are excluded. Responses are recorded using a 5-point Likert scale in which 1 is equal to very unlikely and 5 is equal to very likely. N=112 for all participants in each treatment; N=56 for each colour of participants in each treatment. Error bars show 95% confidence intervals. Stars show the significance of the difference computed from a two-sided t-test: *p<0.10, **p<0.05, ***p<0.01.

Tables

Table E.1: Misattribution: Effect of cumulative average share of purple hires on believed probability of the employer being type A (Rounds 2+)

Dependent Variable:	Subjective probability that employer is type A			
	Rounds 2+			
	(1)	(2)	(3)	(4)
Share of hires purple	0.640***	0.668***	0.996***	0.994***
	(0.110)	(0.089)	(0.237)	(0.241)
Session FEs		✓	✓	✓
Individual FEs			✓	✓
Round FEs				✓
Baseline mean	0.458	0.458	0.458	0.458
Observations	1280	1280	1280	1280
R^2	0.159	0.214	0.604	0.608

Notes: Linear fixed effects model estimation of the subjective probability that the employer is type A, the taste-based discriminator, on the cumulative share of previous hires that go to purple candidates. Round 1 is excluded. Sample includes the baseline treatment and first four rounds of *revealed* treatment. Baseline mean is the average subjective probability that the employer is type A based of the whole included sample where the average share of hires going to purple candidates was 58.5%. Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.2: Misattribution: Effect of cumulative average share of purple hires on subjective probability of the employer being type A excluding irrational decision makers

Dependent Variable:	Subjective probability that employer is type A							
	All rounds				Rounds 2+			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Share of hires purple	0.630***	0.718***	0.858***	0.864***	0.669***	0.711***	1.049***	1.043***
	(0.077)	(0.065)	(0.102)	(0.123)	(0.104)	(0.084)	(0.232)	(0.241)
Session FEs		✓	✓	✓		✓	✓	✓
Individual FEs			✓	✓			✓	✓
Round FEs				✓				✓
Baseline mean	0.471	0.471	0.471	0.471	0.455	0.455	0.455	0.455
Observations	1479	1479	1479	1479	1232	1232	1232	1232
R^2	0.217	0.260	0.563	0.567	0.174	0.230	0.610	0.615

Notes: Linear fixed effects model estimation of the subjective probability that the employer is type A on the cumulative share of hires that are purple. Participants who are irrational in their application decision are excluded. Baseline mean is the average subjective probability that the employer is type A based of the whole included sample where the average share of hires going to purple candidates was 60.5% (columns 1-4) and 58.5% (columns 5-8). Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.3: Misattribution: Effect of cumulative average share of purple hires on subjective probability of the employer being type A excluding by round pairings

Dependent Variable:	Subjective probability that employer is type A							
	Rounds 1-2		Rounds 3-4		Rounds 5-6		Rounds 7-8	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Share of hires purple	0.698***	0.709***	0.523**	0.565**	3.049**	3.049**	1.632*	1.844**
	(0.072)	(0.069)	(0.190)	(0.196)	(0.960)	(0.987)	(0.720)	(0.603)
Session FEs	✓	✓	✓	✓	✓	✓	✓	✓
Individual Controls		✓		✓		✓		✓
Baseline mean	0.495	0.491	0.477	0.477	0.461	0.461	0.438	0.438
Observations	512	512	512	512	256	256	256	256
R^2	0.342	0.374	0.188	0.234	0.279	0.339	0.333	0.387

Notes: Linear fixed effects model estimation of the subjective probability that the employer is type A on the cumulative share of hires that are purple. Even columns include individual level controls in place of individual fixed effects. These are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Baseline mean is the average subjective probability that the employer is type A based of the whole included sample where the average share of hires going to purple candidates was 65.1% (columns 1-2), 59.5% (columns 3-4), 57.9% (columns 5-6), and 55.9% (columns 7-8). Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.4: Misattribution: Effect of cumulative average share of purple hires on believed probability of the employer being type A, for participants with productivity values in the set $[0, 90)$

Dependent Variable:	Believed probability that employer is type A							
	All rounds				Rounds 2+			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Share of hires purple	0.578***	0.739***	0.901***	0.879***	0.565***	0.686***	0.957**	0.942**
	(0.070)	(0.066)	(0.140)	(0.125)	(0.108)	(0.082)	(0.344)	(0.350)
Session FEs		✓	✓	✓		✓	✓	✓
Individual FEs			✓	✓			✓	✓
Round FEs				✓				✓
Baseline mean	0.466	0.466	0.466	0.466	0.448	0.448	0.448	0.448
Observations	622	622	622	622	512	512	512	512
R^2	0.193	0.246	0.671	0.675	0.130	0.197	0.705	0.710

Notes: Linear fixed effects model estimation of the subjective probability that the employer is type A on the cumulative share of hires that are purple. Participants who are irrational in their application decision are excluded. Sample is restricted to observations at which the cumulative average share of purple hires was at least 50% and in which a productivity value in the set $[0, 90)$ is drawn. Baseline mean is the average subjective probability that the employer is type A based of the whole included sample where the average share of hires going to purple candidates was 60.4% (columns 1-4), and 58.3% (columns 5-8). Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.5: Misattribution perpetuates discrimination: Effect of subjective probability of the employer being type A, by rounds

Dependent Variable:	Application decision							
	Rounds 3-8				Rounds 2-4			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Subj. prob. of type A	-1.211*	-1.732**	-1.702**	-1.766**	-0.511	-1.062	-1.382	-1.444
(Previous round)	(0.726)	(0.713)	(0.834)	(0.816)	(1.005)	(1.115)	(1.518)	(1.512)
Purple	-1.544***	-1.729***	-1.677***	-1.717***	-1.009*	-1.227**	-1.316**	-1.300**
	(0.545)	(0.599)	(0.634)	(0.606)	(0.564)	(0.583)	(0.651)	(0.652)
Purple × Subj. prob. of type	2.632**	2.956**	3.123**	3.279**	1.930	2.267	2.723*	2.754*
A (Previous round)	(1.217)	(1.330)	(1.532)	(1.491)	(1.359)	(1.447)	(1.616)	(1.608)
Effect of belief on purple	1.454**	1.267*	1.461	1.551*	1.419	1.204	1.341	1.310
Session FEs		✓	✓	✓		✓	✓	✓
Individual Controls			✓	✓			✓	✓
Round FEs				✓				✓
Baseline mean	0.557	0.557	0.557	0.557	0.544	0.544	0.544	0.544
Observations	332	332	332	332	193	193	192	192
R^2	0.139	0.179	0.221	0.236	0.175	0.234	0.291	0.293

Notes: Logistic fixed effects models of the propensity of application on the subjective probability that the employer is type A. The independent variable is interacted with an indicator of whether the participant is purple colour to show the heterogeneous effects on application for each colour. Participants who are irrational in their application decision are excluded. All specifications include a quadratic control for productivity level and a linear control for the cumulative share of hires going to purple workers from the previous round (the hiring history observable when the participant makes their application decision). Sample is restricted to participants who draw a round productivity value in the interval $[0, 90)$. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.6: Effect of misattribution on fairness perceptions at endline survey

Dependent Variable:	Fairness view of hiring process at endline (5-point Likert scale, 1 = Very Unfair and 5 = Very Fair)					
	Baseline			Baseline + Revealed		
	(1)	(2)	(3)	(4)	(5)	(6)
Subj. prob. of type A	-0.946*	-1.268***	-1.022**	-0.658**	-0.961**	-0.797**
	(0.369)	(0.201)	(0.355)	(0.279)	(0.305)	(0.268)
Purple	-0.348	-0.356	-0.306	0.026	0.025	0.040
	(0.554)	(0.541)	(0.583)	(0.187)	(0.190)	(0.204)
Purple × Subj. prob.	0.464	0.517	0.334	-0.111	-0.074	-0.247
	(0.867)	(0.897)	(0.972)	(0.477)	(0.519)	(0.518)
of type A	(0.867)	(0.897)	(0.972)	(0.481)	(0.519)	(0.518)
Effect of belief on purple	-0.481	-0.752	-0.689	-0.769	-1.035*	-1.044*
Session FEs		✓	✓		✓	✓
Individual Controls			✓			✓
Observations	96	96	96	176	176	176
R^2	0.071	0.157	0.213	0.154	0.198	0.233

Notes: Linear fixed effects model estimation of the effect of the belief of employer being type A, the taste-based discriminator, on respondent's fairness views of the hiring procedure after all rounds. Sample is restricted to participants who observe a hiring history in round 8 of at least 50% of the hires going to purple candidates. All specifications include controls for the cumulative average share of hires that are purple by round 4. Specifications 4-6 include a treatment dummy. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.7: Effect of misattribution on fairness perceptions across all treatments

Dependent Variable:	Fairness view of hiring process (5-point Likert scale, 1 = Very Unfair and 5 = Very Fair)					
	Midline			Endline		
	(1)	(2)	(3)	(4)	(5)	(6)
Purple	-0.257	-0.250	-0.257	0.012	0.023	-0.008
	(0.163)	(0.173)	(0.204)	(0.150)	(0.155)	(0.177)
Subj. prob. of type A	-0.018***	-0.017***	-0.018***	-0.012***	-0.013***	-0.013***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Purple × Subj. prob. of type A	0.007*	0.007	0.007	-0.002	-0.002	-0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)
Constant	5.092***	5.036***	4.624***	4.810***	5.006***	4.528***
	(0.247)	(0.394)	(0.699)	(0.552)	(0.495)	(1.050)
Effect of belief on purple	-0.011***	-0.011**	-0.011**	-0.014***	-0.015***	-0.015***
Session FEs		✓	✓		✓	✓
Individual Controls			✓			✓
Observations	288	288	288	272	272	272
R^2	0.163	0.177	0.201	0.173	0.204	0.226

Notes: Linear fixed effects model estimation of the effect of the belief of employer being type A, the taste-based discriminator, on respondent's fairness views of the hiring procedure after round 4. Sample is restricted to participants who observe a hiring history in round 4 (columns 1-3) or round 8 (columns 4-6) of at least 50% of the hires going to purple candidates. All specifications include controls for the cumulative average share of hires that are purple by round 4/8 and treatment dummy variables. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.8: Effect of misattribution on ex-ante perceived efficacy of a temporary quota

Dependent Variable:	Ex-ante perceived efficacy of a quota					
	Baseline			Baseline + Revealed		
	(1)	(2)	(3)	(4)	(5)	(6)
Subj. prob. of type A	0.566	0.620	0.525	0.654	0.745	0.661
	(0.642)	(0.721)	(0.656)	(0.467)	(0.613)	(0.643)
Purple	-0.086	-0.050	-0.202	0.039	0.050	-0.020
	(0.355)	(0.375)	(0.264)	(0.245)	(0.253)	(0.240)
Purple × Subj. Prob of type A	0.032	-0.050	0.091	-0.162	-0.215	-0.179
	(0.625)	(0.699)	(0.537)	(0.485)	(0.516)	(0.475)
Effect of belief on purple	0.599	0.570	0.616	0.492*	0.530	0.481
Session FEs		✓	✓		✓	✓
Individual Controls			✓			✓
Baseline mean	3.156	3.156	3.156	3.176	3.176	3.176
Observations	128	128	128	256	256	256
R^2	0.062	0.088	0.158	0.053	0.080	0.113

Notes: Linear fixed effects model estimation of the effect of the subjective probability of employer being type A, the taste-based discriminator, on respondent's subjective probability that a temporary quota would increase the share of green hires in subsequent rounds. Columns 1-3 use only baseline data while columns 4-6 include the *revealed* treatment adding a dummy control for the treatment condition. All specifications include controls for the cumulative average share of hires that are purple by round 4. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.9: Effect of misattribution and ex-ante perceived efficacy on demand for temporary affirmative action

Dependent Variable:	Bid for one period of quota					
	Baseline			Baseline + Revealed		
	(1)	(2)	(3)	(4)	(5)	(6)
Subj. prob. of type A	0.872	0.419	0.093	0.138	-0.465	-0.132
	(2.147)	(2.200)	(1.954)	(1.922)	(1.905)	(1.538)
Purple	3.756	4.434	3.862	1.806	2.322	1.718
	(3.146)	(3.663)	(3.660)	(2.017)	(2.078)	(1.962)
Purple × Subj. Prob	-8.199	-8.696	-8.232	-6.660**	-6.822**	-6.682**
	(4.941)	(4.697)	(4.508)	(2.624)	(2.473)	(2.316)
Ex-ante efficacy	1.969***	2.108***	1.895**	1.810***	1.874***	1.759***
	(0.513)	(0.602)	(0.564)	(0.378)	(0.407)	(0.383)
Purple × Ex-ante efficacy	-0.708	-0.847	-0.939	-0.420	-0.569	-0.546
	(0.732)	(0.780)	(0.789)	(0.510)	(0.546)	(0.545)
Effect of belief on purple	-7.327*	-8.278**	-8.139**	-6.522***	-7.287***	-6.814***
Effect of ex-ante eff. on purple	1.261	1.261	0.956	1.390**	1.305**	1.213**
Session FEs		✓	✓		✓	✓
Individual Controls			✓			✓
Baseline mean	6.898	6.898	6.898	6.867	6.867	6.867
Observations	128	128	128	256	256	256
R^2	0.062	0.088	0.158	0.053	0.080	0.113

Notes: Linear fixed effects model estimation of the effect of the subjective probability of employer being type A, the taste-based discriminator, and on the ex-ante perceived efficacy of a quota on respondent's bid for one period of affirmative action. Columns 1-3 use only baseline data while columns 4-6 include the *revealed* treatment adding a dummy control for the treatment condition. All specifications include controls for the cumulative average share of hires that are purple by round 4. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.10: Effect of misattribution on demand for temporary affirmative action (extensive margin)

Dependent Variable:	Probability of making non-zero bid for quota					
	Baseline			Baseline + Revealed		
	(1)	(2)	(3)	(4)	(5)	(6)
Purple	-0.032	0.006	-0.265	0.001	0.003	-0.192
	(0.822)	(0.822)	(0.805)	(0.371)	(0.383)	(0.384)
Subj. prob. of type A	1.471	1.343	1.827	1.211	1.428	1.647
	(1.168)	(1.372)	(1.457)	(0.941)	(1.074)	(1.087)
Purple × Subj. prob.	-2.052	-2.285	-2.528	-2.119*	-2.251*	-2.424**
	(1.721)	(1.767)	(1.664)	(1.098)	(1.160)	(1.065)
Avg. marginal effect on green	0.243	0.212	0.246	0.230	0.258	0.263
	(0.194)	(0.202)	(0.181)	(0.179)	(0.186)	(0.171)
Avg. marginal effect on purple	-0.138	-0.207	-0.135	-0.205	-0.175	-0.153
	(0.195)	(0.172)	(0.188)	(0.205)	(0.110)	(0.097)
Session FEs		✓	✓		✓	✓
Individual Controls			✓			✓
Baseline mean	0.688	0.688	0.688	0.688	0.688	0.688
Observations	128	128	128	256	256	256
Psuedo R^2	0.050	0.100	0.201	0.026	0.070	0.140

Notes: Logit fixed effects model estimation of the effect of the subjective probability of employer being type A, the taste-based discriminator, on propensity of an individual to make a non-zero bid for one period of affirmative action. Columns 1-3 use only baseline data while columns 4-6 include the *revealed* treatment adding a dummy control for the treatment condition. All specifications include controls for the cumulative average share of hires that are purple by round 4. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Average marginal effect is the estimated change in probability from holding a subjective probability of 0% to 100% on the propensity to make a non-zero bid. Baseline mean is the share of participants making a non-zero bid at a mean subjective probability of the employer being type A of 44.4% (columns 1-3) and 22.2% (columns 4-6) where the subjective probability is hard-coded to be 0% for participants in the *revealed* treatment. Standard errors for coefficients are clustered at the society level. Standard error for the average marginal effect computed using the Delta-method. *p<0.10, **p<0.05, ***p<0.01

Table E.11: Effect of misattribution on demand for temporary affirmative action (intensive margin)

Dependent Variable:	Bid for one period of quota					
	Baseline			Baseline + Revealed		
	(1)	(2)	(3)	(4)	(5)	(6)
Purple	2.568	2.077	0.719	0.910	0.731	-0.026
	(2.876)	(3.001)	(2.710)	(1.012)	(1.104)	(0.855)
Subj. prob. of type A	-1.087	-2.031	-2.537	-1.275	-2.921	-2.690
	(2.575)	(3.345)	(2.684)	(2.077)	(2.832)	(2.278)
Purple × Subj. prob. of type A	-7.197	-6.362	-3.726	-4.642**	-4.434*	-3.006*
	(4.544)	(4.604)	(3.708)	(2.055)	(2.081)	(1.661)
Effect of belief on purple	-8.284*	-8.392*	-6.263*	-5.917**	-7.355**	-5.696**
Session FEs		✓	✓		✓	✓
Individual Controls			✓			✓
Baseline mean	10.034	10.034	10.034	9.989	9.989	9.989
Observations	88	88	88	176	256	176
R^2	0.107	0.190	0.295	0.052	0.135	0.241

Notes: Linear fixed effects model estimation of the effect of the subjective probability of employer being type A, the taste-based discriminator, on participants' bid for one period of affirmative action, conditional on the bid being greater than 0. Columns 1-3 use only baseline data while columns 4-6 include the *revealed* treatment adding a dummy control for the treatment condition. All specifications include controls for the cumulative average share of hires that are purple by round 4. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table E.12: Average guessed maximum productivity of applicants, by round

	Round number							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Subjective max. productivity of								
Green applicants	68.777	69.531	71.961	70.398	70.531	68.117	71.648	74.203
	(3.183)	(3.089)	(3.310)	(3.335)	(5.014)	(5.307)	(5.058)	(4.829)
Purple applicants	74.984	78.941	77.805	75.426	77.422	79.297	79.578	82.016
	(3.105)	(3.233)	(3.231)	(3.295)	(3.295)	(4.644)	(4.926)	(4.941)
Two-sided t-test								
Difference	-6.207***	-9.410***	-5.848**	-5.027	-6.891	-11.180***	-7.930**	-7.813**
	(2.020)	(2.799)	(2.862)	(3.189)	(4.480)	(4.131)	(3.482)	(3.282)
Share predicting pooling strategy	0.293	0.215	0.215	0.258	0.234	0.211	0.234	0.188
Observations	256	256	256	256	128	128	128	128

Notes: Average guesses of the maximum productivity value of green and purple applicants in one's society. Sample is full sample in the pooled baseline data including baseline treatment and first 4 rounds of *revealed* treatment. Standard deviations are displayed in parentheses. Standard errors are displayed in parentheses. Share predicting pooling strategies maps the share of participant-round observations in which the reported maximum productivity of green and purple applicants are equivalent. Stars of the difference correspond to the significance of two-sided t-tests to test the equivalence of guesses about green and purple applicants: *p<0.10, **p<0.05, ***p<0.01.

Table E.13: Average guessed maximum productivity of applicants, in *strategic* treatment

	(1)	(2)	(3)	(4)
Subjective max. productivity of				
Green applicants	54.220	51.558	55.317	57.282
	(1.347)	(1.381)	(1.769)	(3.021)
Purple applicants	58.534	54.948	63.074	64.894
	(1.331)	(1.333)	(1.531)	(2.381)
Two-sided t-test				
Difference	-4.314***	-3.390**	-7.757***	-7.612**
	(1.469)	(1.620)	(1.487)	(2.375)
Share predicting pooling strategy	0.154	0.143	0.160	0.102
Excludes round 1		✓	✓	✓
Average purple hires ≥ 50%			✓	✓
Believe type A > 50%				✓
Observations	1024	896	608	255

Notes: Average guesses of the maximum productivity value of green and purple applicants in one's society. Full sample refers to all participants in the *strategic* treatment arm. Column (2) excludes guesses made in round 1. Column (3) excludes participants who observe a hiring history from the previous round in which less than 50% of hires going to purple candidates. Column (4) includes only those whose previous round's belief of the probability that the employer is type A was above 50%, the Bayesian benchmark. Standard errors are displayed in parentheses. Share predicting pooling strategies maps the share of participant-round observations in which the reported maximum productivity of green and purple applicants are equivalent. Stars of the difference correspond to the significance of two-sided t-tests to test the equivalence of guesses about green and purple applicants: *p<0.10, **p<0.05, ***p<0.01.

Table E.14: Effect of *strategic* treatment information on misattribution

Dependent variable	Subjective probability that employer is type A					
	(1)	(2)	(3)	(4)	(5)	(6)
Share of hires purple (0-100)	56.755***	65.413***	67.436***	67.646***	97.381***	105.463***
	(6.713)	(6.217)	(5.869)	(5.703)	(10.148)	(15.781)
Strategic	-1.360	1.294	1.667	2.164	5.910**	10.077*
	(2.966)	(1.894)	(1.843)	(1.804)	(2.143)	(4.841)
Perceived distance	0.010	0.013	0.011	0.014	0.019	-0.004
	(0.014)	(0.012)	(0.011)	(0.011)	(0.013)	(0.023)
True distance	0.038**	0.031**	0.033**	0.033**	0.022	0.010
	(0.015)	(0.013)	(0.013)	(0.013)	(0.017)	(0.027)
Strategic × True distance	-0.052**	-0.042***	-0.042***	-0.042**	-0.037*	-0.029
	(0.020)	(0.015)	(0.015)	(0.015)	(0.020)	(0.055)
Effect of True distance in *Strategic*	-0.015	-0.011	-0.009	-0.009	-0.015	-0.019
Session FEs		✓	✓	✓	✓	✓
Individual controls			✓	✓	✓	✓
Round FEs				✓	✓	✓
Excl. share of purple hires <50%					✓	✓
True distance > perceived distance						✓
Baseline mean	45.495	45.495	45.495	45.495	50.521	52.396
Observations	2512	2512	2512	2512	1824	902
R^2	0.170	0.208	0.237	0.242	0.190	0.204

Notes: Linear fixed effects models estimating the effect of the *strategic* treatment on the subjective probability of the employer being type A controlling for current hiring histories and guesses of maximum productivity values of green and purple applicants. The dependent variable is multiplied by 100 to represent a percentage to aid the readability of coefficients. True distance refers to the difference between the true maximum productivity values of purple applicants and that of green applicants. Participants in society-rounds for which no applicants from a specific group applied are excluded (N=48). The sample excludes participants in the *revealed* treatment after round 4. Specifications 4, 5 and 6 include individual controls for age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Specifications 5 and 6 excludes participants facing a hiring history in which the share of purple hires is less than 50%. Specification 6 incudes only individuals who believe the difference between the maximum productivity of purple and green applicants is less than the true difference. Baseline mean is the average subjective probability of the employer being type A where the average share of hires going to purple candidates is 57.9% (columns 1-4), 66.8% (column 5) and 69.1% (column 6). Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01.

F Deviations from the Pre-Analysis Plan

The analyses reported in this paper follow the pre-registered plan closely with the following exceptions.

Econometric specification The pre-analysis plan specified the use of individual and session *random* effects in the primary regression analyses. Upon inspection of the data, I find that stable individual-level heterogeneity in beliefs and application behaviour is substantial, motivating a switch to *fixed* effects throughout the main specifications. Fixed effects are a more conservative and robust approach in the presence of individual-level confounds, and I retain the pre-specified random effects specification as a robustness check. Standard errors are clustered at the society level in all specifications, as pre-specified. I present a random effects model which tests for the presence of misattribution in Table F.1. The conclusion from the fixed effects specifications are robust to the use of random effects instead.

The pre-analysis plan remained agnostic on the correct method to model a potential time trend. I pre-specified four possibilities: fixed effects, linear, quadratic and logarithmic. The main specification uses fixed effects; I present all four cases in Table F.2 for the presence of misattribution and in Table F.3 for the direct effect of misattribution. The exact time trend used does not affect the identification of misattribution nor its direct consequences on application behaviour.

Testing for misattribution The key deviation from the specified analysis is to accommodate for the fact that not every society experienced discrimination that could be misattributed. This was due to the random draws in the experiment. Therefore, I defined a discriminatory outcome as one which favours purple workers: in which the cumulative average share of hires going to purple is at least two thirds. In Figure F.1, I demonstrate that misattribution is robust to any definition of discrimination where the minimum share of hires going to purple candidates is above 50%.

In the pre-analysis plan, I specified to test whether participants' justifications for their subjective probabilities were more representative of those that overestimate the probability of the employer being type A compared to the Bayesian benchmark. I registered a prior that statistical deduction would be most common among misattibrutors. This is confirmed by the distribution of responses in Figure F.2. I also registered the priors that the random and hunch justifications were most frequent among subjective probabilities close to 50% while the hope justification was most common among subjective probabilities close to 0%. Figure F.3 provides evidence for the latter while demonstrating low frequencies of random and hunch justifications among participants with high final subjective probabilities.

Model for direct effects of misattribution The pre-analysis plan pre-specified to plot application decisions for participants over a number of belief bins. I separate participants into 3 bins of subjective probabilities of the employer being type A and present the average application rate for each colour in the pooled baseline data in Figure F.4. While not significant, the application rate declines for green workers and increases for purple workers as reported probabilities increase.

I pre-specified a linear regression model with random effects for the analysis of application decisions. I pre-specified two separate specifications for each colour. However, restricting the sample to participants in rounds for which they drew productivity values in the range $[0, 90)$ resulted in a smaller sample than anticipated and thus power concerns merited a combined analysis. Given the marginal analysis which was executed, a non-linear estimation was more appropriate to identify the non-linear effect of subjective probability on application propensity at different productivity levels.

Period groupings for robustness. The pre-analysis plan specified that results would be reported both excluding the first two rounds (to account for learning) and excluding the last four rounds (to account for fatigue). Upon examining the data, I find that the primary source of atypical behaviour is concentrated in round 1, consistent with a learning effect at the onset of the experiment. I therefore report results both for the full sample and excluding round 1 only.

Testing for cause of misattribution The pre-analysis plan outlined a three-stage mediation analysis to test whether the effect of the strategic treatment on application behaviour operates through its effect on beliefs of employer type. Given that the treatment did not significantly affect the subjective probability of the employer being type A, a full three-stage mediation analysis would be uninformative and is therefore not reported. Instead, I provide the two-stage analysis of treatment effects on beliefs and application strategies separately, which is sufficient to characterise the null result for the hypothesis that participants misattribute due to cursedness.

Figures

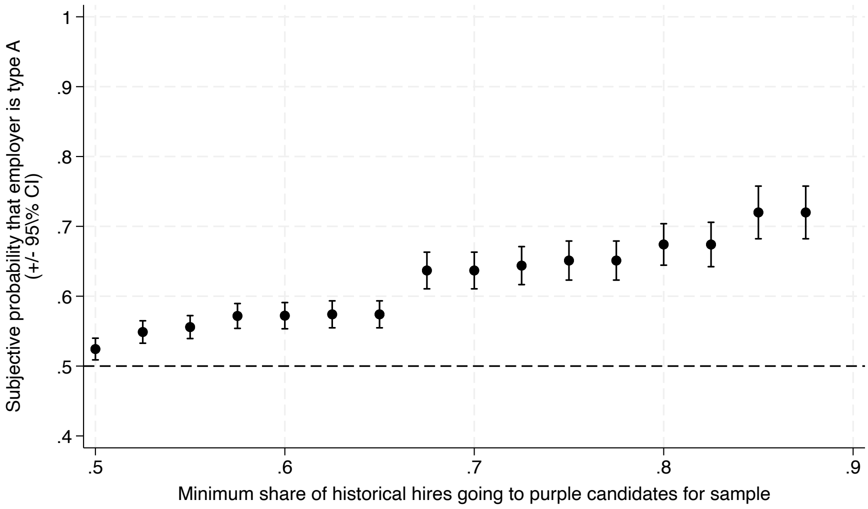


Figure F.1: Average subjective probability that the employer is type A at different threshold values of hiring histories

Note: This graph shows the average subjective probability that the employer is type A (the taste-based discriminator) among individuals who observe a hiring history of at least some share being purpose. Error bars show 95% confidence intervals.

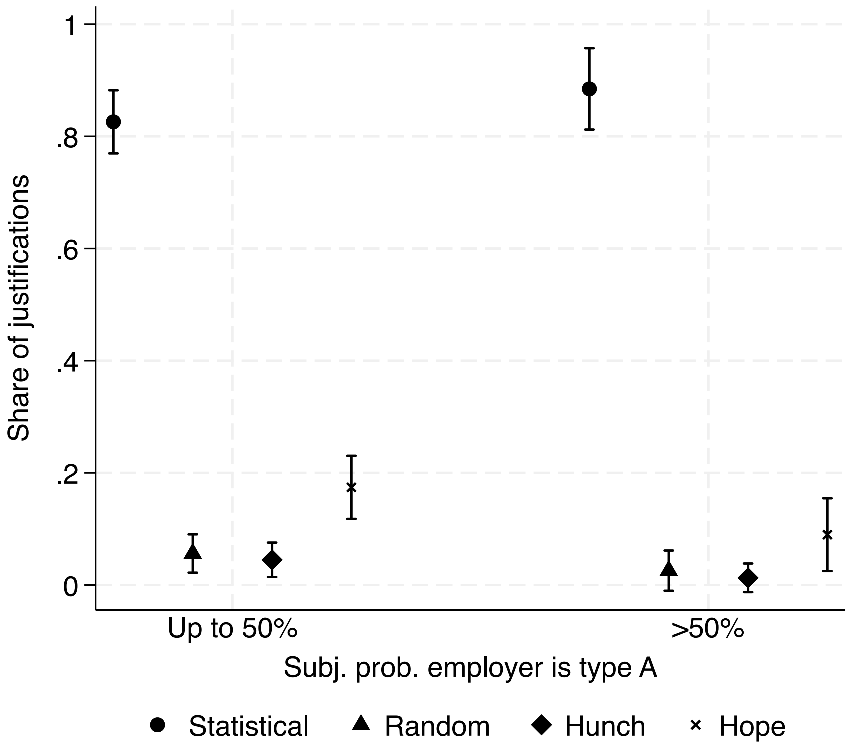


Figure F.2: Share of participants who use each justification category by whether their final subjective probability of the employer being type A is above 50%

Note: This graph shows the share of participants who justified their final subjective probability of the employer being type A using (i) statistical reasoning referring to the hiring histories, (ii) randomisation in that they were not certain of the true type, (iii) a hunch or intuition meaning they followed some heuristic, or (iv) hope in that they expressed they believed it was a certain type because it would bring them utility. 173 of the 256 responses were successfully coded by Claude Sonnet 4.6 using the following prompt 'I would like you to analyse a data file. In this, I have observations of individuals coded by their id. I want you to analyse the text in variable 'explanation'. Specifically, I would like you to code whether the text corresponds to one or multiple or none of the following criteria and add a variable for each which is 1 or 0. The criteria are 'Use of hiring data and/or statistical or probabilistic reasoning'. 'Explicit mention of randomising their choice or uncertainty', 'Explicit mention of following a hunch, intuition or no idea', 'Explicit mention of hope or any forward looking expression'. I hand-coded the remaining 83. Error bars show 95% confidence intervals.

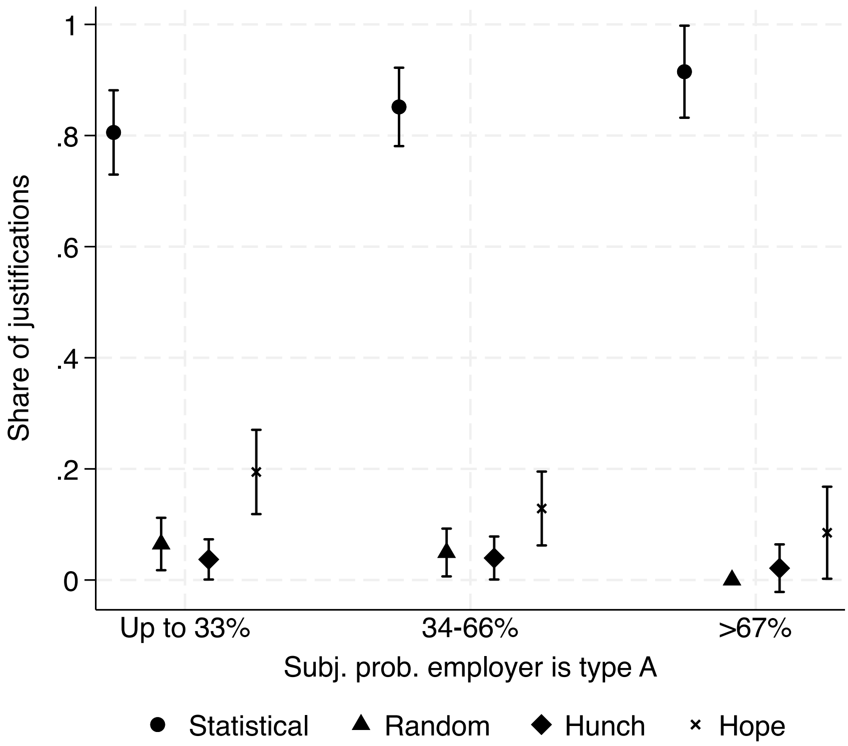


Figure F.3: Share of participants who use each justification category by their final subjective probability of the employer being type A

Note: This graph shows the share of participants who justified their final subjective probability of the employer being type A using (i) statistical reasoning referring to the hiring histories, (ii) randomisation in that they were not certain of the true type, (iii) a hunch or intuition meaning they followed some heuristic, or (iv) hope in that they expressed they believed it was a certain type because it would bring them utility. 173 of the 256 responses were successfully coded by Claude Sonnet 4.6 using the following prompt 'I would like you to analyse a data file. In this, I have observations of individuals coded by their id. I want you to analyse the text in variable 'explanation'. Specifically, I would like you to code whether the text corresponds to one or multiple or none of the following criteria and add a variable for each which is 1 or 0. The criteria are 'Use of hiring data and/or statistical or probabilistic reasoning'. 'Explicit mention of randomising their choice or uncertainty', 'Explicit mention of following a hunch, intuition or no idea', 'Explicit mention of hope or any forward looking expression'. I hand-coded the remaining 83. Error bars show 95% confidence intervals.

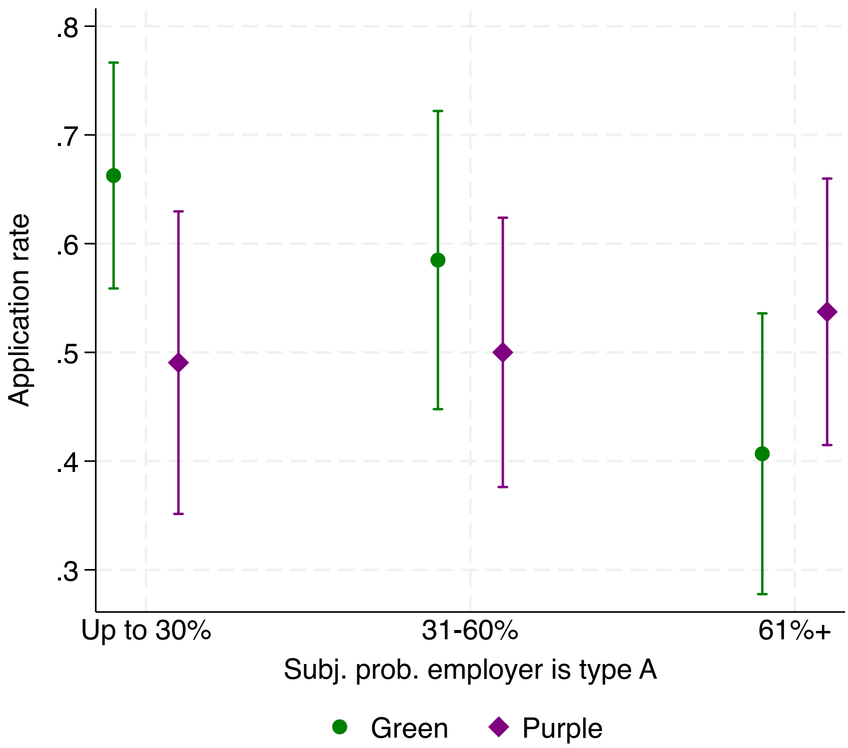


Figure F.4: Share of participants with each bin of subjective probability applying to a job

Note: This graph shows the share of participants who applied to a job conditional on their reported probability that the employer is type A from the previous round. Sample includes baseline treatment and first four rounds of *revealed* treatment. Error bars show 95% confidence intervals.

Tables

Table F.1: Misattribution: Effect of Cumulative Share of Purple Hires on Believed Probability of Employer Being Type A

Dependent Variable:	Believed probability that employer is type A							
	All Rounds				Rounds 2+			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Share of hires purple	0.607***	0.710***	0.804***	0.804***	0.640***	0.734***	0.899***	0.899***
	(0.080)	(0.071)	(0.040)	(0.040)	(0.110)	(0.118)	(0.064)	(0.064)
Constant	0.117**	0.072	0.004	0.004	0.118	0.094	0.038	0.038
	(0.049)	(0.050)	(0.044)	(0.044)	(0.068)	(0.071)	(0.056)	(0.056)
Estimator	OLS	RE	RE	RE	OLS	RE	RE	RE
Individual REs		✓	✓	✓		✓	✓	✓
Session REs			✓	✓			✓	✓
Period REs				✓				✓
Observations	1,536	1,536	1,536	1,536	1,280	1,280	1,280	1,280
Participants	256	256	256	256	256	256	256	256
Societies	12	12	12	12	12	12	12	12
R^2	0.200				0.159			
Wald statistic ($\chi^2(1)$)		100.18	397.24	397.24		38.84	195.74	195.74

Notes: Models (1) and (5) are OLS with standard errors clustered by society. Models (2) and (6) use individual-level random effects (`xtreg, re`). Models (3)–(4) and (7)–(8) use nested random effects (session > society > individual), estimated via REML (`mixed`); model (4) and (8) additionally include crossed period random effects. The independent variable is the normalized cumulative average share of hires who are purple. Standard errors clustered at the society level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table F.2: Misattribution: Effect of cumulative average share of purple hires on believed probability of the employer being type A (different time trends)

Dependent Variable:	Subjective probability that employer is type A							
	All rounds				Rounds 2+			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Share of hires purple	0.854***	0.854***	0.859***	0.858***	0.994***	1.000***	0.997***	1.001***
	(0.120)	(0.111)	(0.120)	(0.118)	(0.241)	(0.242)	(0.242)	(0.243)
Session FEs	✓	✓	✓	✓	✓	✓	✓	✓
Individual FEs	✓	✓	✓	✓	✓	✓	✓	✓
Time trend	FEs	Linear	Quad.	Log.	FEs	Linear	Quad.	Log.
Baseline mean	0.474	0.474	0.474	0.474	0.458	0.458	0.458	0.458
Observations	1536	1536	1536	1536	1280	1280	1280	1280
R^2	0.565	0.561	0.561	0.561	0.608	0.604	0.604	0.604

Notes: Linear fixed effects model estimation of the cumulative share of hires that are purple on the outcome variable of subjective probability that the employer is type A, the taste-based discriminator. Columns 1 and 5 use the round fixed effects, columns 2 and 6 substitute these fixed effects for a linear time trend. Columns 3 and 7 include a quadratic time trend and columns 4 and 8 include a logarithmic time trend instead. Independent variable is then the normalised cumulative average share of purple hires. Baseline mean is the average subjective probability that the employer is type A based of the whole included sample where the average share of hires going to purple candidates was 60.5% (columns 1-4), and 58.5% (columns 5-8). Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

Table F.3: Misattribution perpetuates discrimination: Effect of subjective probability of the employer being type A

Dependent Variable:	Application decision			
	(1)	(2)	(3)	(4)
Subj. prob of type A	-2.068**	-2.003**	-1.998**	-2.004**
(Previous round)	(0.849)	(0.848)	(0.851)	(0.850)
Purple	-1.504***	-1.486**	-1.486**	-1.485**
	(0.554)	(0.581)	(0.582)	(0.582)
Purple × Subj. prob of type A (Previous round)	3.402**	3.306**	3.313**	3.296**
	(1.357)	(1.400)	(1.398)	(1.398)
Effect of belief on purple	1.334**	1.303*	1.315*	1.291*
Session FEs	✓	✓	✓	✓
Individual Controls	✓	✓	✓	✓
Time trend	FEs	Linear	Quad.	Log.
Baseline	0.539	0.539	0.539	0.539
Observations	381	381	381	381
R^2	0.241	0.231	0.231	0.230

Notes: Logistic fixed effects models of the propensity of application on the subjective probability that the employer is type A. The independent variable is interacted with an indicator of whether the participant is purple colour to show the heterogeneous effects on application for each colour. Column 1 uses the round fixed effects, column 2 substitutes the fixed effects for a linear time trend. Column 3 includes a quadratic time trend and column 4 include a logarithmic time trend instead. All specifications include a quadratic control for productivity level and a linear control for the cumulative share of hires going to purple workers from the previous round (the hiring history observable when the participant makes their application decision). Sample is restricted to participants who draw a round productivity value in the interval $[0, 90)$. Individual level controls are age, gender, education level (proxied by whether one is studying for or has achieved a degree higher than a Bachelor's), student status, 5-point Likert scales on economic and social political leaning, and risk preferences as proxied for by a Gneezy and Potters (1997)-style investment game. Standard errors displayed below the coefficient estimates are clustered at the society level. *p<0.10, **p<0.05, ***p<0.01

G Experiment Screenshots

Instructions

Welcome.

Please enter your seat number below and then click on 'Start Study' when told to do so.

Your seat number

Start Study

Welcome & Consent

Thank you for taking part in this study. The study will last approximately 1 hour and you will receive a participation fee of £8.00. During the study, you will make a number of decisions in a number of rounds. At the end of the session, there will be a survey. You are able to earn bonus payments in all parts of the study (including the survey) which will be paid to you on top of your participation fee.

During the instructions, please do not click on anything until told to do so

We kindly ask you not to discuss the study with anyone else inside or outside this room.

Consent

All data collected in this study is anonymised. At any point in the study, you may opt out. If you wish to do so, please raise your hand and I will come to you.

All information in this experiment referring to other participants is true and relates to real human participants in this room. The decisions which you make in this study may impact your own bonus payment and that of other participants. Your identity throughout the experiment will always remain anonymous to others.

There are no known risks to this study and you are not required to provide any personally identifying information.

This study has received ethical approval from the Norwegian School of Economics' Institutional Review Board. Should you have any further questions regarding this study, please contact ethan.oleary@nhh.no.

Should you agree to these terms, please indicate your consent by checking the box below. If you disagree, you may withdraw now by raising your hand.

☐ **I agree to participating in this study and understand how the data I provide will be used.**

Next

Overview of the study

In this study, you will play 8 rounds of a hiring game.

In each round, you will make 3 different kinds of decisions, which each earn you tokens (1 token = £0.03).

At the conclusion of the study, one round will be randomly selected for each of the three decisions. These selections are made independently. Your final bonus payment will be calculated based on your earnings from the selected round corresponding to each decision.

For example, your bonus may consist of, say, your earnings from decision 1 in round 7, decision 2 in round 3, and decision 3 in round 5.

Therefore, it is in your best interest to treat every decision in every round as if you will be paid for it.

There are 4 pages of study instructions and 1 page outlining the decisions that you will make in each round. The instructions will take approximately 15-20 minutes to cover. If, at any point, you have a question, please raise your hand and I will come over and answer your question. **Pay close attention to the instructions as you will need to pass a test to complete the study.**

Next

Your Society and Your Colour (Instructions Page 1/4)

Your society

For the experiment, you will be matched with 15 other participants in this room to form a **society**. You will interact **only** with the participants in your society for the entirety of today's session.

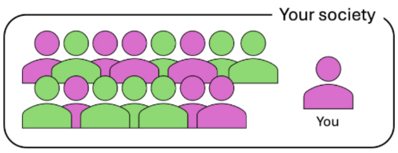
Your colour

You have been randomly assigned a colour indicating your **group identity** in the society. Your colour will remain fixed for the entirety of the study.

In your society there will be 8 Green and 8 Purple participants (including you).

Colours are purely for the purposes of this study and have no connection to any real world trait.

Your colour is PURPLE



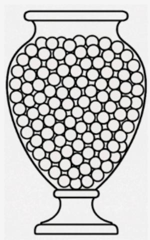
Next

Productivity Values (Instructions Page 2/4)

At the beginning of each round, each participant is given a **productivity value**. Your productivity value and also the productivity values drawn by other members of your society will influence your bonus payments from this study. Below we describe how this value is allocated.

1 The Productivity Urn ▼

Productivity Urn
1000 balls
Average ball value: 50



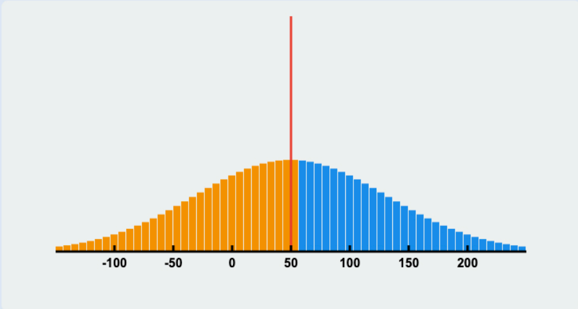
- Imagine an urn called the **Productivity Urn** which contains **1,000 balls**, each with a **whole** number written on them (e.g. 1, 5, 10).
- The **average ball number** in the Productivity Urn is 50: some numbers are negative, some are very large but **most are close to 50**.
- **Half the balls** have numbers **below** 50: the other half are at or above 50.

In each round, each participant independently draws **one ball** from this urn. The number on that ball is their **productivity value**.

You will only see your own productivity value and you should keep this information private.

2 Explore the Productivity Distribution

Use this tool to explore the probability of getting different ***productivity values****.*
In ***orange*** *is the probability of getting a productivity* ***less than or equal*** *to the value.*
In ***blue*** *is the probability of getting a productivity* ***more than or equal*** *to the value.*
Slide the dot along to explore different values.



Slide to explore values:

Productivity: 50

Probability of getting a productivity value ≤ 50:
50%

Probability of getting a productivity value ≥ 50:
50%

3 Your Productivity Value for Round 1

You have drawn a productivity value of 74 for round 1.

4 Your Application Decision (Decision 1)

In each round, your first decision is to choose whether to **apply for a 'job'**. You do not need to do any work for the job and your payment is not tied to any future effort.

If you choose to **apply** for a job, then you will join the applicant pool together with others in your 16-person society who also choose to apply in that round. An **employer** will choose **3 individuals from the applicant pool to hire**.

- **If you are hired**, you will earn **90 tokens** for this decision.
- **If you are not hired**, you will earn **0 tokens** for this decision.

If you choose **not to apply** for a job, your earning for this decision will be **your productivity value** in tokens. For example, if, in some round, your productivity value is 55 and you choose not to apply, then you will earn 55 tokens for decision 1. If your productivity value is instead -10, then you will earn -10 tokens for this decision if you choose not to apply.

Summary of your earnings for decision 1.

Do not apply	**Apply**	
	Hired	**Not hired**
Productivity value in tokens	90 tokens	0 tokens

On the next page, we will explain how the employer chooses which applicants to hire.

Next

How Signals Work (Instructions Page 3/4)

Your productivity value relates, in some way, to how profitable you are to the employer. The employer would like to hire the applicants they **believe to be the most profitable** to them.

When deciding who to hire, the **employer** does not see participants' true productivity values. Instead, they observe a **signal**. Your **signal** is determined by your productivity value and a number drawn from another urn - the Noise Urn. Read more about this below.

1 **Productivity vs Noise Urn** ▶

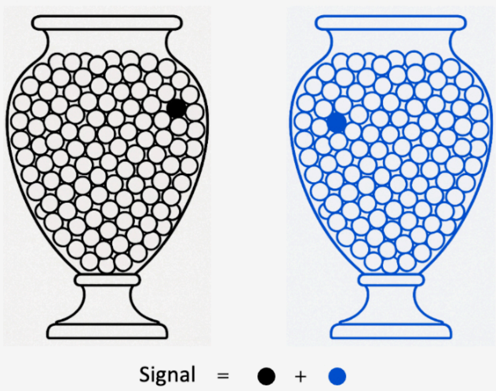
2 Explore your **signal** distribution ▶

3 Key Points Summary ▶

1 **Productivity vs Noise Urn** ▼

If you choose to apply for the job, you will draw a ball from the Noise Urn. You will not see this draw. The number on the ball drawn from the Noise Urn is added to your productivity value to make your **signal**. Each participant who applies will draw their own ball from this urn independently.

Productivity Urn	Noise Urn
1 000 balls	1 000 balls
Average = 50	Average = 0



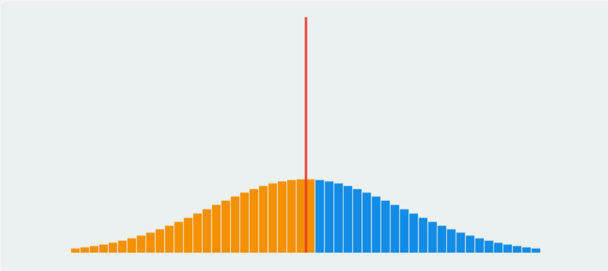
- There are 1000 balls in the **Noise Urn**, each with a number on them.
- The **average ball number** in the Noise Urn is 0: the numbers are **spread around** 0 (some negative, some positive but most are close to 0).
- **Half the balls** have numbers **below** 0: the other half are at or above 0.

As an example, you drew a productivity value of 109. If you draw a ball with the number 14 from the Noise Urn, then the employer will see a signal of 123.

2 Explore your signal distribution

This tool is calibrated specifically to your round 1 productivity value. *Use it to explore the probability of getting different* ***signals****. The signal is your productivity value (109) plus the number on the ball you draw from the noise urn.*

Therefore the probability of you getting some signal is directly related to the probability of drawing a number from the noise urn.



Slide to explore values:

Signal: 109

Probability of getting a signal at or below 109:	Probability of getting a signal at or above 109:
50%	**50%**

3 Key Points Summary

- For each applicant, the employer sees only **group colour** and **signal**.
- Each individual in your society who applies will have a signal centred around the productivity value they drew in that round.
- Your signal may be lower or higher than your productivity value.
- The employer wants to hire applicants it **believes** will have the highest productivity.
- Your true productivity remains hidden from the employer when they make the hiring decision.

On the next page is a short quiz to check your understanding of the preceding instructions. Afterwards, you will learn about the employer and how they interpret signals.

Next

The Employer (Instructions Page 4/4)

The employer uses a **computer robot** to decide which 3 applicants to hire.

Before the study begins, **one of the below robots will be selected** to be used for **all the hiring decisions in your society** for the duration of the study. Robot A is selected with some probability *p* and robot B with some probability *1-p*.

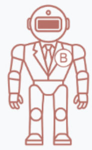
These robots differ in which workers they believe are more profitable and in how they use available information to make decisions.

Robot A

Uses a *systematic adjustment protocol*:

- **Believes Green workers** are less profitable than Purple workers.
- Adjusts Green signals by subtracting 90.
- Adjusts Purple signals by adding 10.
- Hires the 3 applicants with the **highest adjusted signal values**.

Robot B



Uses *advanced statistical modeling*:

- Uses **all available information** (payoffs, instructions, signals of applicants from previous rounds) to predict which workers are most likely to apply from **each colour group**.
- **Estimates applicant's actual productivity** using this prediction together with the applicant's **signal and colour**.
- Hires the 3 applicants with the **highest estimated productivity**.

Important:

1. Robot A or Robot B will be chosen at the beginning of the game.
2. The **same robot** is used to evaluate all applicants from your society in each round of the study.
3. You do not know which robot is chosen.

Next

Your Decisions in each round (last page before rounds begin)

In each round, you will be shown your productivity value for that round, reminded of your colour, and see a graphic which shows you the previous hiring decisions **for your society**.

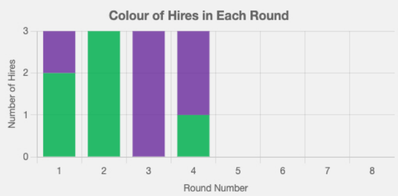
Below is an example as if you are in round 5.

In each round, you need to make three decisions. This graph will update before you make your third decision. Each vertical column represents the hiring decision in a specific round. The most recent round is always the bar to the right of the graph. In the example below, for instance, 1 green worker and 2 purple workers were hired in round 4.

Round 5

Your productivity value: X
Your colour: PURPLE

Decision ?



Below, you will read about each decision that you must make in each round and the corresponding payment rules. You will complete each decision in order on separate pages in each round.

1 **Decision 1: Application Decision** ▶

2 **Decision 2: Who Applies** ▶

3 **Decision 3: Which Robot** ▶

You must wait for all individuals in your society to complete decisions 1 and 2 before moving to decision 3. In the first couple of rounds, it is ok to spend time to get to know the decisions. However, to ensure that others are not held back, we ask that you do not spend longer than 1 minute per decision in the following rounds.

A summary of all of the instructions is printed for you on the rear page of your handout. You can refer to this at any time.

Move to Round 1 Start

1 **Decision 1: Application Decision** ▼

You will decide whether to apply for a job in each round

- **If you apply for a job:** You will earn 90 tokens if you are hired and 0 tokens if you are not hired (only 3 people in your society will be hired in each round).
- **If you do not apply:** You will earn your productivity value in tokens.

2 **Decision 2: Who Applies** ▼

You will guess the **maximum productivity value** among the participants in your society who chose to apply in the current round, separately for

- Green participants.
- Purple participants (excluding you if you apply).

In other words, from all the *other* participants who choose to *apply* in your society this round, please estimate the highest productivity value for the green group and for the purple group.

For each of the two guesses:

- If your guess is within 10 of the correct answer: you earn 50 tokens.
- If it is between 11 and 20 of the correct answer: you earn 25 tokens.
- Otherwise: you earn 0 tokens.

Therefore, in terms of your earnings, it is in your best interest to report your true belief.

Your token balance for decision 2 is the sum of your earnings from each guess.

3 **Decision 3: Which Robot** ▼

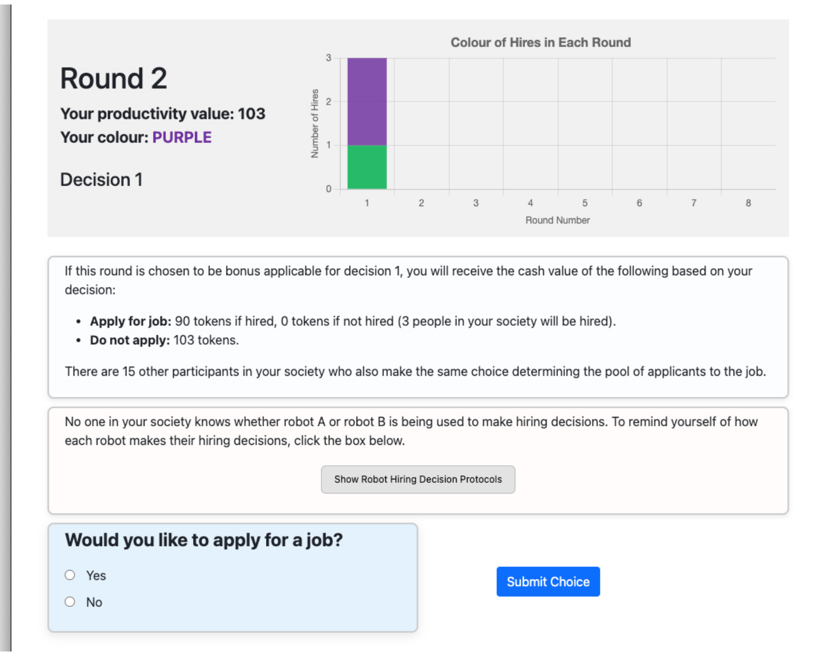
When you make decision 3, the hiring graph will be updated to include the current round. In other words, you will, at this point, be able to observe the colours of participants who were hired in the current round.

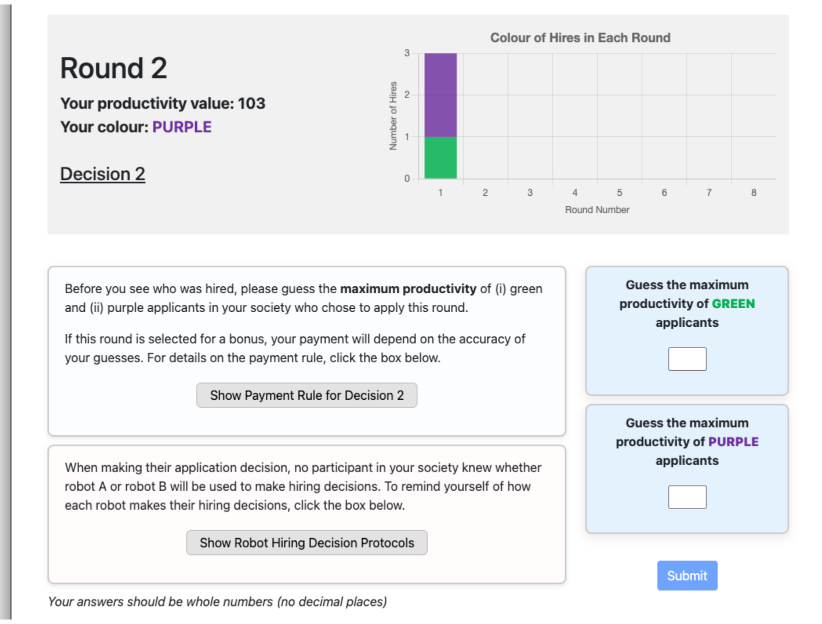
You will guess **which robot is being used** to make employment decisions for your society.

- You will **allocate 100 tokens** between Robot A and Robot B based on **how likely** you think it is that each robot is being used.
- For example if you believe that there is a 60% chance that robot A is being used and a 40% chance that robot B is being used, you should allocate 60 tokens to robot A and 40 to robot B.
- The **payment rule** follows a formula that is proven to ensure that **reporting your true belief gives you the highest expected earnings**.
- Therefore, you should report your true belief of the likelihood that each robot is being used. The more accurately your token allocation reflects your true belief, the higher your expected earnings will be.
- If you wish to see the formula for computing your earnings, click below.

Show Formula for Payment Rule for Decision 3

Decisions


Round 2
Your productivity value: 103
Your colour: PURPLE
Decision 1
Colour of Hires in Each Round
Number of Hires
Round Number
If this round is chosen to be bonus applicable for decision 1, you will receive the cash value of the following based on your decision:
Apply for job: 90 tokens if hired, 0 tokens if not hired (3 people in your society will be hired).
Do not apply: 103 tokens.
There are 15 other participants in your society who also make the same choice determining the pool of applicants to the job.
No one in your society knows whether robot A or robot B is being used to make hiring decisions. To remind yourself of how each robot makes their hiring decisions, click the box below.
Show Robot Hiring Decision Protocols
Would you like to apply for a job?
Yes
No
Submit Choice

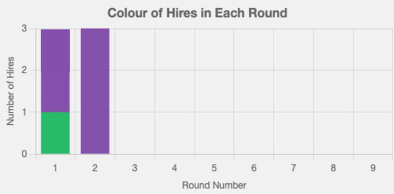

Round 2
Your productivity value: 103
Your colour: PURPLE
Decision 2
Colour of Hires in Each Round
Number of Hires
Round Number
Before you see who was hired, please guess the maximum productivity of (i) green and (ii) purple applicants in your society who chose to apply this round.
If this round is selected for a bonus, your payment will depend on the accuracy of your guesses. For details on the payment rule, click the box below.
Show Payment Rule for Decision 2
When making their application decision, no participant in your society knew whether robot A or robot B will be used to make hiring decisions. To remind yourself of how each robot makes their hiring decisions, click the box below.
Show Robot Hiring Decision Protocols
Guess the maximum productivity of GREEN applicants
Guess the maximum productivity of PURPLE applicants
Submit
Your answers should be whole numbers (no decimal places)

Round 2

Your productivity value: -5
Your colour: GREEN

Decision 3

In this round, the robot in use decided to hire 0 **green** workers and 3 **purple** workers, as is shown.



The robot used to make these hiring decisions was **either Robot A or Robot B**. **You must now guess which robot is in use.** To be reminded of how each robot makes this hiring decisions, click the box below.

Show Robot Hiring Decision Protocols

Below, you must **allocate 100 tokens** between Robot A and Robot B based on **how likely** you think it is that each robot is being used. For details on the payment rule, click the box below.

Show Payment Rule for Decision 3

You will maximise your expected earnings by reporting your true belief.

Please indicate how likely you believe each robot is making the hiring decisions.

(When you enter a value in one box, the other box will update to ensure you allocate all 100 tokens. Text will appear to show you your earnings with your current input values. This text may only update the first few times you input new numbers. If it no longer updates with new numbers, you can ignore the text and continue to input your true belief.)

Robot A Enter 0-100

Robot B Enter 0-100

Midline survey

Interlude Page 1

To what extent do you consider the hiring procedure has been fair thus far?

Very Unfair	Somewhat Unfair	Neither Unfair nor Fair	Somewhat Fair	Very Fair
○	○	○	○	○

Next

Interlude Page 2

In **Round 5 only**, a **quota policy** may be applied. While this policy is in place, the robot is obligated to hire a certain mix of workers. You will first bid how much you would pay to have the policy enacted. **You can bid up to 20 tokens.** Then a decision criteria will determine whether the policy is implemented.

Quota Policy Details

If the quota policy is enacted, a fair coin will be flipped:

- **If it lands on heads:** the robot must hire **2 green** and **1 purple** worker in round 5.
- **If it lands on tails:** the robot must hire **1 green** and **2 purple** workers in round 5.

In all cases, if the policy is enacted, at least one worker of each colour is hired.

Step 1: Submit Your Bid

- Choose a number of tokens to bid from 0 to 20.
- Your bid expresses how much you value having the quota policy.

Step 2: Decision Criteria for Implementation and Payment

- One participant from your society is randomly picked to be the **decision maker**.
- A number from 0 to 20 is drawn randomly, this number is called the **draw**.
- If the draw is less than or equal to the decision maker's bid, the quota policy is enacted in round 5. Otherwise, the quota policy is not enacted.
- If the quota policy is enacted, the decision maker is deducted a number of tokens equal to the draw form their *final bonus payment*.

Example

- Suppose you are randomly chosen to be the decision maker and you bid 12 tokens.
- If the draw is 8, then the policy will be enacted in round 5 and you will be deducted 8 tokens from your final payment.
- If the draw is 15, then the policy is not enacted and no tokens are deducted from your final payment.

It is always in your best interest to bid your true valuation of the policy. Recall that, regardless of whether the policy is enacted, you also have 3 additional rounds to play (rounds 6 to 8) where no policy will be enacted.

How much would you like to bid for the policy to be enacted in round 5?

Enter 0-20 tokens

Submit

Interlude Page 3

How likely do you think it is that implementing the quota in round 5 will increase the share of green hires in later rounds when no quota is in place (rounds 6–8)?

Very unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Very likely
○	○	◉	○	○

Next