(Mis)anticipated discrimination: Pre-Analysis Plan

Ethan O'Leary*

November 14, 2025

This paper investigates how beliefs about the underlying source of discrimination - taste-based or statistical - affect the emergence, persistence and consequences of discriminatory outcomes. Using a representative U.S. survey, I document that individuals systematically overestimate the incidence of taste-based discrimination in society. I argue that such misperceptions can sustain discriminatory outcomes even in the absence of genuine preferential biases. To formalise this mechanism, I develop a theoretical model in which individuals, anticipating discrimination, condition their job application decisions on identity, thereby rationalising statistical discrimination in a set of equilibria. The application and hiring strategies in a discriminatory equilibrium are identical for when employers are information-constrained rational agents versus when they harbour taste-based discrimination, potentially leading workers to misattribute the source of discrimination. The misattribution of the source of discrimination particularly sustains discriminatory outcomes as individuals fail to recognise that their beliefs, not employer animus, drive these patterns, causing them to undervalue deviations from equilibrium application strategies. I hypothesise that one mechanism which may lead to misattribution is cursedness: workers fail to internalise that others' application behaviour is itself identity-contingent leading rational employers to statistically discriminate. In a laboratory experiment, I demonstrate that individuals do misattribute discriminatory outcomes to taste-based motives and this is driven by cursed strategic reasoning. I test for consequences of misattribution on demand for temporary anti-discrimination policies and perceptions of fairness in the labour market.

JEL: J7, D91, J2, C92
Keywords: Anticipated Discrimination, Cursed Equilibrium, Affirmative Action, Application Gap

*FAIR The Choice Lab & Department of Economics, Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway. Email: ethan.oleary@nhh.no.

Experimental set-up and sample

Public. The experiment will run at the CeDEx facility in the University of Nottingham. Across the span of one week (17th November 2025 to 20th November 2025), I will run 12 experimental sessions. In each session, I will run a maximum of 2 treatments. Participants are randomly assigned to one of two societies in each session which indicate the group of participants in the session who will interact and the treatment assignment. Each treatment will be run over 8 societies and each society will consist of 16 participants.[1]

Private. I must pre-register two important parameterisations of the experiment. First, the true value of p, the probability that robot A is used, is 0.01 or 1%. Prior to running the first session, I drew 24 random integers between 1 and 100. I specified that if any of these integers were 1, then I would have to run the corresponding number of sessions where the true robot in use was A. All the drawn integers were above 1 so all sessions therefore run with robot B making hiring decisions.

Second, the initial belief of robot B regarding who applies is crucial to the path that societies are set on. The non-discriminatory equilibrium of the model, given other parameters is that participants should apply if their productivity value is no greater than 38.5. In the discriminatory equilibrium, green participants apply if and only if their productivity value is 1 or less while purple participants apply if and only if their productivity value is 58 or less. I set the prior belief of robot B to then be the midpoint along the *satiation* curve between these two equilibria. This is that the prior belief of maximum productivity value of green applicants is 19.75 and that of purple applicants is 51.5. Simulation data suggested that to achieve sufficient convergence to either of the possible equilibria strategies while permitting a degree of learning the task before equilibria are reached, I should set the standard deviation of this prior to be 15. A full visualisation of the simulations is in the appendix.

I Experiment Design

Baseline I will recruit groups of 16 participants to form *societies* who are initially randomly divided into two group-identities defined by their colour: green and purple.[2] They are also assigned a *productivity value* which is loosely assigned from a normal distribution of mean 50 and standard deviation 81.85. [3][4] Participants are given a choice on how they wish to earn their bonus payment. They may take a bonus equal to their *productivity value* in tokens or may apply for one of three *jobs* which pay a fixed bonus payment of 90 tokens. Participants who unsuccessfully apply are given a bonus payment equal to 0 tokens.

[1] In the case that I do not manage to recruit 32 participants to a session, I will run a session with societies consisting of 14 participants.

[2] Green and purple are used to separate colours from any real affiliations such as politics or gender.

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

[4] To simplify matters for the participants, I inform then that there are 1000 balls in an urn, each with an integer written on them. Half of the balls have a number below 50 and half at or above 50. Negative numbers exist. Participants will be presented with a tool for which they can slide a bar along and see how many of the balls in the urn have a number higher and lower than their selected value. This tool will be visible at all times.

The job application evaluation process is as follows. A single employer wishes to hire 3 *workers* from the applicant pool. The employer's profits are positively influenced by the *productivity value* of the workers they hire. However, the employer is unable to observe the true *productivity value* of applicants at the point of evaluation. Instead, they observe two features of each applicant: their group identity (green or purple) and a noisy but unbiased signal of the applicants' true *productivity value*. The signal is the sum of the productivity value and a stochastic noise element which has a commonly known normal distribution of mean 0 and standard deviation of 81.85.[5]

The employer uses a robot to conduct the evaluation and hiring process on their behalf. Before the experiment, a random draw determines which one of two robot types is employed throughout the whole procedure. Robot A systematically adjusts signals according to the colour of the applicant: it adds 10 to signals from purple applicants and subtracts 90 from signals of green applicants, and then hires the 3 applicants with the highest adjusted signal values. Robot B hires the applicants that it predicts have the highest *productivity* given the estimated conditional distributions of *productivity values* among the likely applicants from each colour computed using the information and incentives given to applicants such as the distribution of types and the colours of previous hires, as well as the distribution of signals in previous rounds. Evidently, robot A practises a naïve taste-based discrimination while robot B employs a fair sophisticated inference which could lead to statistical discrimination should application strategies differ by groups. The description of each robot is left purposefully vague to allow for subsequent treatments to fill in these gaps. In all sessions, robot B is always employed.

Participants are first asked whether they wish to apply for a *job* or not. They are then incentivised to guess the maximum *productivity value* among the applicants from each group identity.[6] After the hiring decisions have been made and the participants can observe the colour of those hired in the current and previous rounds, they are then incentivised to guess which robot type has been employed. I give participants 100 tokens and ask them to distribute these between the two robot types, employing the quadratic scoring rule for incentivisation.

Participants repeat this process for twelve rounds. Across all rounds, the participant's group identity and the robot type employed are fixed. In each round, each participant draws a new *productivity value*. After all rounds, participants are paid their bonus token balance from one of the rounds chosen at random at an exchange rate of 100 tokens = £2.00.

Treatments Treatments are assigned on the society level and are summarised in table 1. Each session contains two societies each with a randomly drawn treatment. The baseline treatment allows me to test for the *mis*attribution of discrimination. I employ two further treatments which allow me to test for the causes and consequences of such respectively.

In the treatment *AVG*, I investigate why individuals misattribute discrimination to taste-

[5]Participants are informed that for each ball in the first urn, there is a second urn which consists of 1000 balls, each with an integer centred around the first ball. Similar explanations are presented as per the first urn. Moreover, a sliding tool is provided again. Given the participants first ball of value x, participants can use a second tool to slide along to see the how many balls in their second urn are at or below a certain number.

[6]A binary scoring rule is employed wherein participants receive 50 tokens if they are within 10 of the true value, 25 tokens if they are within 20 of the true value and nothing otherwise.

based sources by relieving participants of the strategic uncertainty in applications. Specifically, I am using this treatment to test whether *cursedness* leads to misattribution and thus allows discrimination to persist. I hypothesise that this uncertainty leads to misperceptions of the underlying mechanics of the discrimination. In this treatment condition, I inform participants of the true maximum ability of applicants of each coloured identity after they predict such and before they predict the robot used. This broadens the scope of the information structure I give to participants to include both the realised outcomes of the game but also the strategies played by others. In this treatment, I aim to convey to participants that *if* they believe that robot B may discriminate when average abilities among applicants differ, then such discrimination may also be occurring.

To test for the consequences of *mis*attribution, I employ the treatment, *INF*, wherein, I inform participants after round 6 that they have been interacting with robot B and therefore, that the discrimination thus far is statistical in nature. After this round in all treatment arms, I measure two attributes and compare individual responses to the same measures conducted in the baseline treatment. First, I measure individual willingness to pay (WTP) for one-round of affirmative action (AA) wherein the employer is forced to hire at least one worker of each group. Under the policy, a coin is flipped to determine a colour. The side the coin lands on dictates which colour must constitute 2 of the hires, and which colour must constitute one of the hires. This decision is incentivised by the following procedure: participants are free to state a number between 1 and 50: a random participant is chosen and a number x is drawn between 1 and 50. If the number is above the participant's stated number, then affirmative action is not enacted in the following round. If the number picked is below the number stated by the participant then affirmative action will occur in the next period and the participant is deducted x tokens from their final payment.[7]

The second measure establishes participants' perception of the fairness of the situation they are in. I employ this measure directly after the willingness to pay elicitation but before results are finalised. I ask participants to rank, on a 5-point Likert scale, to what extent they consider the hiring process thus far to be fair.

Post-experiment survey One may be concerned that my measure of WTP for AA may not appropriately capture the intended preferences. Thus, I follow up this measure with a post-experiment survey. In this survey, I ask participants how effective a one-period quota would have increased the share of green hires and applicants in later rounds.

I also employ a risk assessment task as per Gneezy and Potters (1997): individuals are given 100 tokens and asked how many of these coins they wish to invest in a lottery which pays 2.5 times the investment with a probability of one third and nothing otherwise.

[7]This incentivisation method is inspired by the Becker-DeGroot-Marschak mechanism (1964), however, I take away the strategic uncertainty inherent in the allocation mechanism. Thus, I reduce the incentive to overestimate true valuations via mechanisms like the winner's curse.

Section	Baseline	INF	AVG
Instructions	Identical		
Round 1-6	Baseline		+ info on max p values of applicants at D3
Midline	Baseline	Reveal true robot	Baseline
Rounds 7-12	As R1-6	+ know true robot (no D3)	As R1-6
Endline	Identical		

Table 1: Explanation of treatment conditions throughout the experimental timeline.

Hypotheses and Data

Hypotheses

I have constructed a theoretical model which derives the following hypotheses except hypothesis four which is conjecture. This model and the derived hypotheses together with proofs are attached to the appendix of this document. The pre-specified primary hypotheses are the following:

1. Participants misattribute the source of discrimination.
2. Misattribution perpetuates discrimination as it leads to a convergence towards separating application strategies.
3. Misattribution is lower when the participants are informed of the true application strategies of other workers.
4. Individuals who believe that they are facing a taste-based discriminator have (i) a lower willingness to pay for one round of a policy, and (ii) a heightened sense of unfairness of procedures, than those who believe that they are facing a statistical discriminator.

Additionally, I am interested in the following secondary hypotheses:

1. Individuals misestimate the true application behaviour of others: overestimating the degree to which green and purple workers differ in their application behaviour.
2. Misattribution is stable across periods with limited information provision.

Data Structure

My data is constructed in the following way. For each individual i of society s in session j, I have multiple observations over periods t. To account for learning, I will report results

separately both excluding and including the first 3 rounds. Moreover, I will also report results excluding the last six rounds to account for potential fatigue.

In addition to demographic data which I collect, a participant's colour, $c_i \in \{\mathcal{P}, \mathcal{G}\}$ is fixed for the duration of the experiment. Each participant is matched to a society which is assigned a treatment $T_i \in \{1, 2, 3\}$ where $T_i = 1$ is the baseline, $T_i = 2$ is the information treatment in which participants are informed of the true robot being used at the midline and $T_i = 3$ is the treatment in which, participants are informed of the true maximum productivity value of applicants after making their guesses.

In each round, t, I collect the following variables from each participant:

- Productivity value draw, q_{ist}.
- Application decision, $A_{ist} \in \{0, 1\}$.
- Belief of max. productivity of (other) green applicants, $\hat{\mathcal{G}}_{ist}$.
- Belief of max. productivity of (other) purple applicants, $\hat{\mathcal{P}}_{ist}$.
- Believed probability of robot A being used, $\hat{\mathcal{R}}_{ist}$.

At the midline, I also collect information on beliefs of the fairness of the hiring process and on preferences for affirmative action. For each participant, I collect the following:

- Perception of the fairness of the procedure up to and including round 6 (5 point Likert scale), f_{is}^6.
- Willingness to pay for one round of an ex-ante equal opportunities quota: $\mathcal{Q}_{is} \in [0, 20]$.
- Belief of probability that a one-round quota will affect hiring outcomes in further rounds (5 point Likert scale), $\hat{\iota}_{is}^6$.

At a random point during the study, I elicit participants' current emotional status by asking them to what extent (on a 5-point Likert scale) do they feel the following:

- Upset
- Calm
- Bored
- Motivated
- Happy
- Anxious
- Tired
- Discouraged
- Annoyed

I collect a bank of data at the end-line survey. For each participant, I collect the following in the end-line:

- The participant's written explanation for why they believe a certain robot is or is not more likely to have been making the hiring decisions. As from the pilot, I will use GPT-4o to code the explanations under the following possible criteria which act as separate binary variables.
  - Use of the hiring data and/or statistical or probabilistic reasoning, ρ_{is}.
  - Explicit mention of randomising choice and uncertainty, r_{is}.
  - Explicit mention of following a hunch, intuition or no idea, k_{is}
  - Explicit mention of hope or any forward looking expression, h_{is}
  - Uncategorised[8], u_{is}
- Perception of the fairness of the procedure in the entire experiment (5 point Likert scale), f_{is}^{12}.
- Belief of the average believed probability that robot A is being used by members of society s in round 1, $\tilde{\mathcal{R}}_{is}^{1}$.
- Belief of the average believed probability that robot A is being used by members of society s in round 12, $\tilde{\mathcal{R}}_{is}^{12}$.
- Belief of probability that a one-round quota will affect hiring outcomes in further rounds (5 point Likert scale), $\hat{\iota}_{is}^{12}$.
- Belief of probability that a one-round quota will affect the application pool in further rounds (5 point Likert scale), $\hat{\alpha}_{is}^{12}$.

Finally, in the end-line, I also collect a bank of individual-level control variables. In the next section, these will be represented in the vector Z_{is}:

- Risk preferences, proxied by the investment in a risky lottery à la Gneezy and Potters (1997).[9]
- Age.
- Gender identity.
- Highest level of education.
- Working status.

[8] I leave the GPT-4o to ascertain whether those explanations that are uncategorised can be separated by another criteria.

[9] Participants are given 100 tokens and asked how many they wish to invest in a lottery paying 1.5 times the investment with a 1/3 chance and nothing with a 2/3 chance. Their investment in the lottery is considered a measure of their risk taking preferences.

- Political party support.[10]

A summary of all the above described variables is in the appendix.

Primary outcome variables

The primary outcome variables for regression analysis are the following for each hypothesis:

- **Hypothesis 1:** Belief of robot A in use, $\hat{\mathcal{R}}_{ist}$.
- **Hypothesis 2:** Application rate as a proxy for the continuation of discrimination A_{ist}.
- **Hypotheses 3 :** Both of the above.
- **Hypothesis 4(i):** Bid for a one-round quota policy Q_{is}.
- **Hypothesis 4(ii):** Self-reported fairness perceptions f^6_{is}.

Secondary outcome variables

- Hiring rates of each group.
- Justification for belief. Used to classify participants into types of thinkers. Pre-specify 4 types and will use GPT to explore possibility for further categories/sub-categories: (i) use of statistical/probabilistic thinking, (ii) randomisation and/or uncertainty, (iii) following a hunch/intuition, and (iv) hopeful or forward looking.
- End-line measures of beliefs of fairness and quota efficacy in affecting (i) hiring outcomes, and (ii) pool of applicants, in future rounds without quota.
- Emotional status during experiment based on 5-item Likert scale ranking feeling of upset, calmness, boredom, motivation, happiness, anxiety, tiredness, discouragement, annoyance.

First Analyses and Visualisations

Exclusion Criteria

Due to the nature of the experiment: a coordination game, it is not possible to entirely exclude the inference of a participant, even if I exclude their individual data.

Participants need to pass the comprehension checks in the instruction phase in order to pass through to the main experiment. Therefore non-comprehension should be infeasible.

I can identify if participants are *saboteurs* by the rationality of their application decisions. I can confidently state that failing to act by a rational set of application strategies is a signal of sabotaging and not a lack of understanding since it is a matter which is clarified explicitly in the comprehension test. That is, I define two rationality principles.

[10] Given the setting in the United Kingdom, I ask participants to state which party they would vote for if a general election were to be held in the following week. I offer support for the top 10 parties in the United Kingdom. Participants are also able to state 'Other' or 'cannot/will not vote'.

1. A participant should always apply to the job if their productivity value is 0 or less.

2. A participant should never apply to the job if their productivity value is 90 or more.

If an individual violates one of these rationality principles during their 12 rounds, then I label them as saboteurs. I will run a supplementary analysis where I exclude these saboteurs in the appendix.

Checks

To describe the data, I first report on the balance of observable characteristics (i) between treatments and (ii) between colours within treatments. The randomisation of treatment to within a specific session should abate any influence of session-timing effects although I will always include session random effects in the regression analyses.

Next, I describe the random draws in the experiment. In each round, individuals are drawn a productivity value and noise values which should be independent of session, colour and round. Therefore, I will map histograms of productivity draws and noise draws by session, colour and round. A more structural test of independence is the χ^2 test of difference in distributions by each factor. Therefore, I will report results from this test.

Treatment effects - Visuals and Averages

I am interested first in describing the key dependent variables: application rates, beliefs of other strategies and beliefs of robot usage. These variables are captured in each round in each treatment. the exception is for treatment 2 in which I do not capture beliefs of robot usage after round 6.

Application rates. Application strategies are contingent on colour. Therefore, I stratify this visualisation by the colour of the participant. A first visualisation is to plot the share of participants applying in each round by colour and treatment. On average, since purple participants in a non-discriminatory equilibrium have a higher threshold for application, the share of purple participants applying should exceed the share of green participants applying. Next, I will plot the average productivity value of **applicants** for each round, colour, and treatment. For similar reasons, if participants are in the discriminatory equilibrium, then this should be revealed by this visualisation. I will employ appropriate formal tests of the difference of the average share of participants of each colour in each treatment applying and the average productivity values in these groups. Finally, I intend to graph a histogram of the average application rate of participants by bins of productivity values by colours.

Beliefs of others' strategies. In each round, participants are asked to report their belief of the **maximum** productivity of green and purple applicants. This is elicited using an open text box that accepts only integers. I will plot the average beliefs of each between treatments and rounds and test for mean differences between the two beliefs. I will also compare these beliefs to the true strategies to ask if individuals correctly anticipate the strategies of other participants.

Belief of robot usage. The belief relating to the robot in use is important to test the hypotheses on misattribution. Before testing such misattribution, I will map the average number of tokens, $\bar{\mathcal{R}}_t$ for each round t and treatment. Here, I compare the belief to the Bayesian benchmark of 50%.

In each round, there are 6 possible combinations of hiring that may occur. I will also split the average belief by the hiring combination that occurs by round. Since beliefs are constantly updated throughout the experiment, I will map the belief in each round compared to the cumulated average share of purple and green hires in previous rounds as a proxy for the *biasedness* of the hiring history. This will require a structural break at round 6 to account for the potential quota.

Hiring rates. Finally, I will check for the hiring rates which are observed. While this is not a primary dependent variable of interest, it is important to the formation of beliefs throughout the experiment. In each round there are 6 possible hiring outcomes representing combinations of green and purple hires. I will plot the average number of hires for each group in each round by each treatment. Next, I will focus on the share of combinations, plotting for each treatment the proportion of societies which experience each hiring outcome in each treatment.

Hypothesis 1

Main test. Hypothesis 1 predicts that individuals misattribute the source of discrimination to robot A (taste-based discrimination) when the true source of discrimination is robot B (statistical discrimination). My main test for hypothesis 1 pools the first 6 rounds of treatments 1 and 2 together and tests for whether the average belief of probability that robot A is being used is above the Bayesian benchmark of 50.

Exploration. To explore this hypothesis, I am interested in a number of potential heterogeneous effects. First, I will explore whether one type of reasoning is more likely to be representative of misattributors. As such, I will stratify the participants into their coded justification method from the end-line survey. My prior is that individuals who use data and/or statistical/probabilistic deduction are most inclined to misattribute the source of the discrimination. Random selection and hunch are most likely to be close to the benchmark, while those following hope are most likely to have a belief in favour in robot B.

Further, I am interested in exploring the stability of misattribution. I will graph a round-by-round average belief and employ a formal test on the average belief per round, comparing this to the Bayesian benchmark and between rounds. To utilise controls, I will employ a regression using individual and session random effects. The exact modelling of the time trend is subject to the data collected and given that there is no comparable experiment to base an assumption of time trends on, pre-specifying a functional form may not be appropriate. I will, however, state that forms I consider to be likely are a linear time trend, individual dummy variables for time periods, quadratic, and logarithmic variations employing multiple hypothesis testing adjustment where appropriate. This test will be run for each treatment on the specified clustering of periods in the Data Structure section.

Hypothesis 2

Main test. Hypothesis 2 proposes that individuals who misattribute discrimination are more likely to end up in the discriminatory equilibrium. To test for this effect, I first stratify the data by colour and treatment. I discard observations from this dataset which have productivity values that are not in the feasible set within which there is no clear dominated application strategy. The feasible set is therefore any productivity value $q_{ist} \in [0, 90]$.

I will normalise the belief $\hat{\mathcal{R}}_{ist}$ to $\hat{R}_{ist} = (\hat{\mathcal{R}}_{ist} - 50)/50$ and stratify the normalised beliefs into at least 3 bins of different levels. The composition and cutoffs of these bins are to be confirmed once the data has been collected and the distribution confirmed.

I will then plot application decisions for participants in each bin over some sub-range of productivity values in the feasible, by colour. I can then formally test the distributional equivalence of productivity values of applicants of each colour for each belief bin. This will inform on whether the beliefs that individuals have over which robot is in use affects the application propensity and therefore discrimination. This test relies on variation of misattribution among the baseline and thus, I will also pool all treatments together for this test.

Exploration. To include controls, I will estimate a linear regression model on the application decision for each colour separately employing the appropriate random effects outlined in the previous section. In this regression, I will employ the previous round's belief of robot usage on the application decision controlling for the current productivity of the individual in the application round. I will estimate the regression for the first two pooled treatments and treatment 3 separately and then pool all treatments together in an additional specification.[11]

Hypothesis 2 implies a prior that the marginal effect of belief on robot A usage on application propensity is negative for green workers and positive for purple workers. Robustness analyses will discretise the belief of robot usage to a dummy variable indicating an individual over-attributing to robot A (an individual who has a belief sufficiently above the Bayesian benchmark).

Hypotheses 3

In this hypothesis, I conduct between treatment analysis. I first compare the average beliefs, $\hat{\mathcal{R}}$, between treatment 3 and the pooled baseline treatment (treatments 1 and 2 are identical in the first 6 rounds) in the specified period groupings to the Bayesian benchmark. This is a sign of whether there is misattribution in each treatment separately.

To compare misattribution intensity between treatments, I will compare the pooled baseline to treatment 3. Indication that the treatment leads to lower misattribution is if the average belief is greater in the pooled baseline than in the comparison treatment. To test for the stability of these differences, I will also (i) plot the per round average beliefs by treatment and (ii) augment the adopted regression specification from hypothesis 1 to include treatment dummy variables.

[11] The intention here is that since treatments may affect beliefs and that application behaviour is hypothesised to be affected only through beliefs, beliefs should capture all treatment effects.

To test for the effect of treatments on application behaviour, I must test for a difference in application strategies between participants in each treatment of each colour. To do this, I will employ a χ^2 test for the equivalent distribution of productivity values of green/purple applicants in the baseline to those in the comparison treatment. To explore further, I will test whether the average productivity value of green/purple applicants is the same between the pooled baseline and comparison treatment groups in the specified period groupings. If the comparison treatment affects application behaviour and thus discrimination, one would expect that treatment 3 increases (decreases) the average productivity value of green (purple) applicants.

Mediation analysis for the Cursed Equilibrium. To test for treatment effects on the main outcome variables through the desired channel, I will employ a mediation analysis which is described below for the estimation of hypothesis 3. This hypothesis is testing for a cursed equilibrium: that individuals misestimate the strategy of others in response to certain information, and that this is causing misattribution and accordingly perpetual discrimination. Initially, this can be demonstrated by observing the difference in beliefs of the maximum productivity values across treatments. If hypothesis 3 is to be true, then I should observe that beliefs of true application strategies proxied by $\hat{\mathcal{G}}_{ist}$ and $\hat{\mathcal{P}}_{ist}$ are closer to the true strategies in treatment 3 than in the pooled baseline. Hypothesis 3 predicts that treatment 3 affects these beliefs which in turn affect the beliefs of robot usage and application behaviour. To do this, I will employ a 2 stage regression analysis of the effect of treatment on belief of robot use instrumenting for beliefs of application strategies.

To ascertain whether the cursed equilibrium affects the perpetuity or stability of discrimination, I will extend this framework to include a third stage wherein I predict robot beliefs using the above second stage and substitute this in to the regression outlined for hypothesis 2.

Hypothesis 4: Responses to Midline information

To test the final set of hypotheses on the consequences of misattribution, I modify the previously used random effects model. Since dependent variables in these hypotheses and not repeated measures, only elicited at the mid-line and final survey, I will employ only session random effects and the vector of individual controls collected in Z_{is}, in lieu of individual effects. To test this hypothesis, I form two sets of comparisons, the first is between treatments. I principally compare treatments 1 and 2 where the belief differences in the robot used are predicted to be most stark. I will first employ a t-test of mean comparisons between the primary set of dependent variables $\{Q_{is}, f^6is\}$. Following this, I will estimate the regression model to incorporate controls, with a dummy variable indicating treatment 2. Robustness analyses test the above with the secondary dependent variables, namely the belief that the quota would've affected hiring outcomes, applicant pools, and the fairness perceptions of the hiring procedure at endline.

The second set of comparisons uses data from all treatments and includes the round 6 belief as an independent variable in place of the treatment dummy. I will then estimate the above comparing the primary and secondary dependent variables between different belief levels.

Variable descriptions

Variable name	Not.	Domain	Timing	Description
Colour	c_{is}	$\{\mathcal{P}, \mathcal{G}\}$	Before r1	Random draw ind-level
Treatment	T_{is}	{1,2,3,4}	Before r1	Random draw society-level
Productivity value (pv)	q_{ist}	$\mathbb{R}$	Every round	Random draw ind-round level
Application	A_{ist}	$\{0, 1\}$	Every round	Choice variable
Belief of green max pv	$\hat{\mathcal{G}}_{ist}$	$\mathbb{R}$	Every round	Choice variable
Belief of purple max pv	$\hat{\mathcal{P}}_{ist}$	$\mathbb{R}$	Every round	Choice variable
Belief of Robot A in use	$\hat{\mathcal{R}}_{ist}$	[0,100]	Every round†	Choice variable
Fairness to r7	f_{is}^{6}	$\{1, 2, 3, 4, 5\}$	Midline	Likert scale self-reported view
WtP for quota in r7	Q_{is}	$[0, 20]$	Midline	Choice variable
Belief of quota affecting hiring outcomes in r7	$\hat{\iota}_{is}^{6}$	$\{1, 2, 3, 4, 5\}$	Midline	Likert scale self-reported view
Use of data/ statistical/ probabilistic deduction	ρ_{is}	$\{0, 1\}$	End-line	Binary indicator from text
Randomly selected robot from uncertainty	r_{is}	$\{0, 1\}$	End-line	Binary indicator from text
Follow hunch/intuition/noidea	k_{is}	$\{0, 1\}$	End-line	Binary indicator from text
Follow hope	h_{is}	$\{0, 1\}$	End-line	Binary indicator from text
Uncategorised decision criteria	u_{is}	$\{0, 1\}$	End-line	Binary indicator from text
Fairness of whole experiment	f_{is}^{12}	$\{1, 2, 3, 4, 5\}$	End-line	Likert scale self-reported view
SOB of probability robot is A in r1	$\tilde{\mathcal{R}}_{is}^{1}$	$[0, 100]$	End-line	Choice variable in survey
SOB of probability robot is A in r12	$\tilde{\mathcal{R}}_{is}^{12}$	$[0, 100]$	End-line	Choice variable in survey
Belief that quota affects hiring outcomes in non-quota r	$\hat{\iota}_{is}^{12}$	$\{1, 2, 3, 4, 5\}$	End-line	Likert scale self-reported view
Belief that quota affects applicant pool in non-quota r	$\hat{\alpha}_{is}^{12}$	$\{1, 2, 3, 4, 5\}$	End-line	Likert scale self-reported view

Table 2: Description of variables collected in the experiment. †unless if treatment group is 2, then only until round 6.

References

Bandiera, O., Jalal, A., and Roussille, N. (2025). The illusion of time: Gender gaps in job search and employment. Technical report, National Bureau of Economic Research.

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

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

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

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

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.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *The 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.

Ruebeck, H. (2024). Perceived discrimination at work. *Available at SSRN 4799864*.

A Theoretical Model

The purpose of this section is to theoretically demonstrate that discrimination may persist via self-confirming beliefs; and to derive testable hypotheses regarding the behavioural misperceptions which may drive the persistence of misattribution. I formalise a model of the preceding example of artisans by building on the seminal model of self-confirming discrimination by Coate and Loury (1993). Specifically, I invert their action and type sets. In the framework of Coate and Loury (1993), agents' labour supply is fixed but they may choose to undertake an unobservable costly investment in skill acquisition. In my model, I fix agents' human capital and endogenise labour supply decisions. Therefore, in the model, workers are given an exogenous productivity value from a known distribution and must decide whether to apply for a fixed-wage job or take their productivity value as an outside option payment, leading to a negative selection of workers into job application.[12]

I motivate this modification through two lines of reasoning. First, recent applied literature 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 when beliefs of discrimination in the labour market arise (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). Second, models of endogenous human capital formation are founded in situations where skill investment decisions are unobservable: an assumption that has not held up against time where diplomas and transcripts are increasingly scrutinised and verifiable.

First I demonstrate that under complete information of no taste-based discrimination, multiple equilibria exist in the model representing both non-discriminatory and discriminatory outcomes. Agents coordinate to one equilibrium depending on their prior anticipation of discrimination. Next, I show that when the employer is sophisticated and unbiased, workers may incorrectly believe that they face a biased employer and thus their anticipatory behaviour sustains the same discriminatory equilibrium as in the rational case which leads to misattribution. I demonstrate that misattribution may occur if workers underestimate the degree of Bayesian sophistication of the employer and if workers are cursed: they underestimate the degree to which other workers condition their application strategies on their identity. Both of these misperceptions incentivise strategies which lead to hiring outcomes that do no test the workers' beliefs of the source of discrimination.

Set-up

Consider a unit mass of risk-neutral workers $i \in [0, 1]$, each of whom lives for one period. Workers are given an exogenous type which constitutes an observable group identity $g_i \in \{A, B\}$ and an unobservable level of ability $q_i \in \mathbb{R}$. Worker types are distributed according to a known process: the share of A group identities is n_A and ability has a distribution $F_q(q)$ with density $f_q(q)$. I assume herein that $F_q(q)$ is a normal distribution with mean μ and variance $1/\tau$ where $\tau \geq 0$ is the concentration of ability.

There is a risk-neutral employer who can hire an exogenous ϕ number of workers to

[12] Such a selection is not intended to model the entire labour market but can describe numerous examples such as employed vs self-employed or public vs private sector employment.

conduct a job which pays a wage ω. The employer is given an exogenous and unobservable type which is a tuple (m, ψ) constituting their relative preferences for hiring an individual from group A over an individual from group B denoted by $m = M^A/M^B$ where $M^g > 0$ is the marginal utility to hiring an individual from group g, and their level of sophistication, $\psi \in \{0, 1\}$ which refers to the employer's ability to anticipate the strategy of the workers. I call an employer fair if they hold no group-specific preferences, $m = 1$, and biased if $m \neq 1$. Moreover, I call them sophisticated if they are aware of the workers' application strategies ($\psi = 1$), and naïve if they are not ($\psi = 0$). I define worker i's beliefs of the employer's type by the vector $(\hat{m}_i, \hat{\psi}_i)$.

Each worker i needs to choose an action $\alpha_i \in \{0, 1\}$ of whether to apply for a job with the employer. If they do not apply then they will earn an outside option of q_i. Application to the job is costless but time-consuming meaning that workers cannot take their outside option in addition to applying.[13] If their application is successful then they are hired and earn a payoff of the wage $\omega > 0$; if they are unsuccessful, then this payoff is 0.[14]

For each worker i who applies, the employer decides whether to hire them $h_i \in \{0, 1\}$. The employer observes a signal of each applicant's type: their group identity and a normal signal of their ability, $s_i = q_i + \epsilon_i$ where $\epsilon_i \sim N(0, 1/\eta)$. Assuming that the employer is insured against losses accountable to individual workers, by hiring a worker i from group g, the employer earns revenue $\max\{M^g q_i, 0\}$.

Since the signal is unbiased with full support, the expected return to hiring an applicant is strictly increasing in the signal that they portray. Thus, the employer's strategy will be to hire all individuals from group g who exhibit a signal at least as great as some signal threshold, $\underline{s}_g$. Formally, this strategy can be denoted by $h(\phi, \psi, m, \hat{\bar{q}})$ which maps the labour market condition ϕ, the employer's type and their belief about the strategy of the workers $\hat{\bar{q}}$ to a tuple of signal thresholds $\underline{s} = (\underline{s}_A, \underline{s}_B)$. I say that a hiring strategy is discriminatory if $\underline{s}_A \neq \underline{s}_B$.

Suppose that workers of group g believe that the employer's strategy is to hire all applicants who give some signal at least as great as $\underline{s}_g$. Then a worker will apply to the job if and only if the following condition holds,

$$q_i \leq \omega \times Pr(s \geq \underline{s}_g | q_i).$$

To simplify the proceedings, I make the following assumption.

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

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

Lemma 1: Existence and properties of a cut-off strategy. When assumption 1 holds, there exists some $\bar{q}_g$ such that an individual of group g applies to the job if and only if they have ability $q_i \leq \bar{q}_g$. **Proof.** See Appendix B for all proofs.

[13] The results are robust to including a non-zero application cost.

[14] One can interpret this as an unemployment insurance contingent on activity in the labour market, similar to that employed in the United Kingdom.

Lemma 1 proves that there is one value of productivity for which an individual is indifferent between applying for the job and taking the outside option. Denote $\bar{q}_g$ as the ability at which an individual from group g is indifferent to applying for the job. Then I can formally define a worker's application strategy by the function $\alpha(\phi, \hat{\psi}, \hat{m}, \hat{\underline{s}})$ which maps the labour market conditions, worker belief of employer type and of the employer's strategies, $\hat{s}$ to a tuple of abilities $\bar{q} = (\bar{q}_A, \bar{q}_B)$. Accordingly, given a strategy α, workers from group g will apply to the job only if their ability is no more than $\bar{q}_g$.

In the following equilibrium, beliefs are self-confirming. Additionally, the hiring strategy must also satisfy constrained profit maximisation to be a best response. This implies two conditions. First, if the employer plays the hiring rule $\underline{s}_A$, it must be profit maximising to set the hiring rule $\underline{s}_B$. Let $\pi(h(\phi, \psi, m, \hat{\bar{q}}))$ be the profit the employer earns playing the strategy h conditional on the labour market definition and employer beliefs of worker strategies, $\hat{\bar{q}}$. Second, the strategy must satisfy labour demand according to the capacity constraint, ϕ. Define $\beta^g(h(\phi, \psi, m, \hat{\bar{q}}))$ as the share of workers from group g hired when playing the strategy h given the labour market definition and employer beliefs of worker strategies.

I next consider the sophisticated and unbiased model where $\hat{\psi} = \psi = 1$ and $\hat{m} = m = 1$. Following, I expand the model to allow for incorrect beliefs.

Complete information of a fair and sophisticated employer

I first consider the case where workers and employers are sophisticated meaning that they hold correct beliefs of each others' types and strategies in equilibrium. I define this self-confirming equilibrium below.

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

$$\bar{q} = \alpha(\phi, \psi, m, h(\phi, \psi, m, \bar{q})). \tag{1}$$

and the employer's strategy meets the following conditions:

1. Optimality of hiring strategy. 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,

$$\pi(h(\phi, \psi, m, \bar{q})) \geq \pi(h'(\phi, \psi, m, \bar{q})) \quad \forall h(\phi, \psi, m, \bar{q}) \neq h'(\phi, \psi, m, \bar{q}). \tag{2}$$

2. Satiation of hiring strategy. The labour demand from the formal employer is satisfied,

$$\sum_{g \in \{A,B\}} n_g \beta^g(h(\phi, \psi, m, \bar{q})) = \phi. \tag{3}$$

From lemma 1, for any signal threshold set, there is a unique application strategy. An employer's optimal hiring strategy sets signal thresholds such that, given their beliefs of $\bar{q}_A$ and

$\bar{q}_B$, the expected return from hiring an applicant from group A with signal $\underline{s}_A$ is equal to the expected return from hiring an applicant from group B with signal $\underline{s}_B$. The conditional expectation of the productivity of applicant from group g given a signal s and application strategy $\bar{q}_g$ can be defined as the following

$$\mathbb{E}[q|s; q \leq \bar{q}_g] = \frac{t(\bar{q}_g)u(\bar{q}_g) + \eta s}{t(\bar{q}_g) + \eta}. \tag{4}$$

Where $t(\bar{q}_g)$ ($u(\bar{q}_g)$) is the concentration (mean) of the distribution of ability among candidates given that their ability is no greater than $\bar{q}_g$. Thus, for self-confirming beliefs to be profit maximising, they must satisfy the following condition

$$\chi(\bar{q}_A) = \chi(\bar{q}_B) \quad \text{where} \quad \chi(z) = \frac{z + x(z) + \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(1 - \frac{2z}{\omega}\right)}{t(z) + \eta}. \tag{5}$$

Where $x(z) = t(z)u(z)/\eta$. Trivially, a solution to (5) is the beliefs $\bar{q}_A = \bar{q}_B$. Lemma 2 additionally outlines the conditions for which some $\bar{q}_A$ and $\bar{q}_B$ where $\bar{q}_A \neq \bar{q}_B$ may simultaneously satisfy (5). Given assumption 1, $\chi(\cdot)$ is decreasing along some subset of the domain of feasible application strategies. Therefore, a necessary condition for this is that there is some set of the domain to $\chi(\cdot)$ in which the function is convex.

Lemma 2: It is always the case that $\bar{q}_A = \bar{q}_B$ is a solution to (5) for all $\bar{q}_g \in (0, \omega)$ and $g \in \{A, B\}$. Also, if assumption 1 holds, when $\sqrt{2\pi/\eta} \geq \omega \geq 2\mu > 0$, there exists some set $\tilde{Q} \subset (0, \omega)$ for which $\chi(q)$ is strictly convex when $q \in \tilde{Q}$.

Lemma 2 demonstrates that non-discriminatory beliefs in the feasible set $(0, \omega)$ can always be self-confirming and optimal. Moreover, under a certain condition, there exists some $q, q', q'' \in (0, \omega)$ such that $q \neq q' \neq q''$ such that $\chi(q) = \chi(q') = \chi(q'')$. In other words, beliefs of different application strategies can be simultaneously self-confirming and optimal.

The satiating condition states that hiring strategies induce application strategies which together result in a mass of ϕ workers being hired. Substituting (1) into equilibrium condition 2 and expanding, one gets the following condition

$$\phi = \sum_g n_g \beta^g(h(\phi, \theta, \bar{q})) \equiv \sum_g n_g \int_{-\infty}^{\bar{q}_g} P\left(e \geq \underline{s}_g - q|q\right) f_q(q) dq. \tag{6}$$

When beliefs are self-confirming and non-discriminatory, (6) reduces to the following

$$\phi = \int_{-\infty}^{\bar{q}} P\left(e \geq \bar{q} - q - \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(\frac{2\bar{q}}{\omega} - 1\right)|q\right) f_q(q) dq. \tag{7}$$

If the above holds for some $\bar{q}$ in the feasible set, then it is also compliant with (5) by lemma 2 proving the existence of a non-discriminatory equilibrium.

Proposition 1: Existence of non-discriminatory equilibrium. Suppose that assumption 1 holds and the employer is known to be fair and sophisticated. If $\phi \leq F_q(\omega)$, there exists a pair of non-discriminatory beliefs $\bar{q}_A = \bar{q}_B$ both in $(0, \omega)$ and signal thresholds $\underline{s}_A = \underline{s}_B$

that satisfy the equilibrium conditions.

Since $\underline{s}_g$ is decreasing in $\bar{q}_g$ in equilibrium, then to satisfy (6) under self-confirming discriminatory beliefs, it must be that as $\bar{q}_A$ increases, $\bar{q}_B$ must decrease. Proposition 2 then states that when the employer is sophisticated, discriminatory equilibria may exist.

Proposition 2: Existence of discriminatory equilibria. Suppose that assumption 1 holds and the employer is known to be fair. There exists some closed sets $\mathbb{M} \subset \mathbb{R}$ and $\Phi \subset [0,1]$ such that if $\mu \in \mathbb{M}$ and $\phi \in \Phi$, then there exists some ω satisfying assumption 1 such that some $\bar{q}_A$ and $\bar{q}_B$ both in $(0,\omega)$ where $\bar{q}_A \neq \bar{q}_B$ simultaneously satisfy the equilibrium conditions.

In Figure 1, I demonstrate the existence of these equilibria where groups are equally numerous ($n_A = 0.5$). On each axis is the belief of the application strategy of some group g. On the dashed line labelled *Satiation*, I map the set of application strategies that are best responses to signal thresholds such that together they satisfy labour demand. In the solid lines labelled *Optimality*, I map the set of application strategies that are best responses to signal thresholds which simultaneously maximise unconstrained profits of the firm. Each intersection of these schedules represent an equilibrium. Along the 45 degree line, I denote the non-discriminatory equilibrium (NDE) while off this line, I identify the two (symmetric) discriminatory equilibria (DE1 and DE2). In a discriminatory equilibrium, one group's applicant pool has a higher average productivity value since they expect a lower signal threshold. This application strategy leads the sophisticated fair employer to set different signal thresholds to satisfy labour demand and unconstrained profit maximisation. Thus beliefs of discrimination are self-confirming.

In the following sections, I show that incorrect beliefs held by workers about the employer type can lead to worker and employer strategies that sustain these equilibria and the incorrect beliefs. I show that workers may misattribute the equilibrium statistical discrimination they face to taste-based discrimination if they underestimate the level of sophistication of employers. Finally, I introduce a cursed equilibrium to the model (Eyster and Rabin, 2005). In this equilibrium, each player assumes that all other players play the same mixed application strategy profile corresponding to the average distribution of actions of the favoured identity: players incorrectly predict that other players do not internalise the anticipation of discrimination in their strategies. I show that this leads workers to incorrectly attribute discrimination to taste-based sources and leads to a persistent discriminatory equilibrium.

Misattribution from incorrect beliefs of employer types

Suppose now that the workers believe that the employer is naïve, $\hat{\psi} = 0$. Intuitively, this means that the workers believe that the employers' beliefs of application strategy always satisfy $\bar{q}_A = \bar{q}_B$. This means that workers do not believe group-specific inference will occur. Without loss of generality, hold some belief of prejudice $\hat{m} \in (1,\infty)$. Then workers believe that the signal thresholds they will face will satisfy $\hat{m}\hat{\underline{s}}_A = \hat{\underline{s}}_B$ such that $\phi = n_A\beta^A(\hat{\underline{s}}_A) + (1-n_A)\beta^B(\hat{m}\hat{\underline{s}}_A)$. Thus, one can show how the application strategies depends on belief $\hat{m}$.

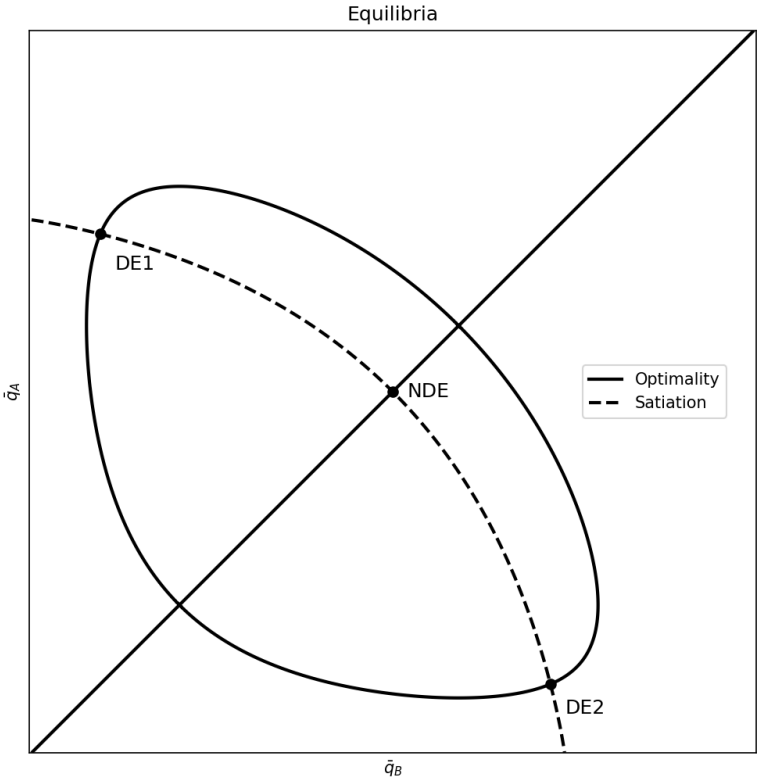

Figure 1: Example of multiple equilibria existing in the rational and fair model. The locus labelled as *Satiation* plots the combination of application strategies that lead to hiring strategies which jointly satisfy labour demand. The loci labelled *Optimality* plots the combination of application strategies that lead to hiring strategies which jointly satisfy unconstrained profit maximisation.

Lemma 3: Say that assumption 1 holds and $\hat{\psi} = 0$. When $\hat{m} = 1$ the application strategies defined by $\bar{q}_A$ and $\bar{q}_B$ are equal and $\bar{q}_A/\bar{q}_B$ is increasing in $\hat{m}$.

Suppose that worker's beliefs are inaccurate and the employer is, in fact, fair and sophisticated. I can define an equilibrium that permits non-consistent beliefs of employer type but consistent beliefs of strategies.

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

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

Given that the employer's strategy meet conditions 1 and 2 in definition 1.

A Bayesian Nash Equilibrium will then be a case where beliefs of naïve, biased and constrained hiring strategies identify a tuple of application strategies that satisfy the PBNE conditions from the previous section.

Proposition 3: Stability of non-discriminatory equilibrium under incorrect belief of employer type. When assumption 1 holds, $\hat{m} = m = 1$ and the workers incorrectly believe the employer to be naïve, then the non-discriminatory equilibrium is the unique equilibrium.

Proposition 3 therefore shows that when workers do not believe in taste-based discrimination but incorrectly believe the employer is naïve, the only equilibrium that is attainable is the non-discriminatory equilibrium.

The following outlines the main theoretical result: that individuals may encounter discrimination of a purely statistical nature but may incorrectly believe that this is caused by a naïve taste-based discriminator and the strategies observed sustain beliefs in discrimination and of the type of discrimination. I refer to this as the misattribution of discrimination.

Proposition 4: Suppose that assumption 1 holds, $m = \psi = 1$ and beliefs are defined as $\hat{\psi} = 0$ and $\hat{m} \neq 1$. Moreover, suppose that the conditions in proposition 2 are satisfied such that the discriminatory equilibria exist for correct beliefs. Then there exists some $\hat{m} \neq 1$ such that the equilibrium conditions in definition 2 hold.

In Figure 2, I include the locus of beliefs of application strategies that are self-confirmed by prejudiced but naïve application strategies at some value $\hat{m}$. The intersection of this locus labelled *Prejudice* with the rational discriminatory equilibrium (DE1) 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 is consistent with their beliefs of discriminatory hiring strategies conditional on their belief of employer type.

Cursed Misattribution from incorrect beliefs of worker strategies

Now suppose that workers hold correct beliefs about the level of sophistication of the employer but underestimate the anticipatory behaviour of other workers and believe that employers are biased, $\hat{m} > 1$. They will believe that they will face signal thresholds satisfying the following profit maximisation condition $\hat{m}\chi(\hat{\bar{q}}_A) = \chi(\hat{\bar{q}}_B)$ where $\hat{\bar{q}}_g$ is their belief of the application strategies of workers from group g and $\hat{\underline{s}}_g$ is their belief of the corresponding hiring strategy.

The assumption of an underestimation of anticiaptory behaviour of workers means that $|\bar{q}_A - \bar{q}_B| > |\hat{\bar{q}}_A - \hat{\bar{q}}_B|$. I assume an extreme level of cursedness where workers incorrectly believe that all other workers do not condition their application strategies to potential discrimination. In this case, it is straightforward to show that they will coordinate to the identical discriminatory equilibrium as above and also misattribute thie source of thid discrimination.

Proposition 5: Cursed Equilibrium. Suppose that assumption 1 holds and $\hat{\psi} = 1$ and $\hat{m} > 1$. Additionally suppose that workers incorrectly believe that $\bar{q}_A = \bar{q}_B$, then there exists

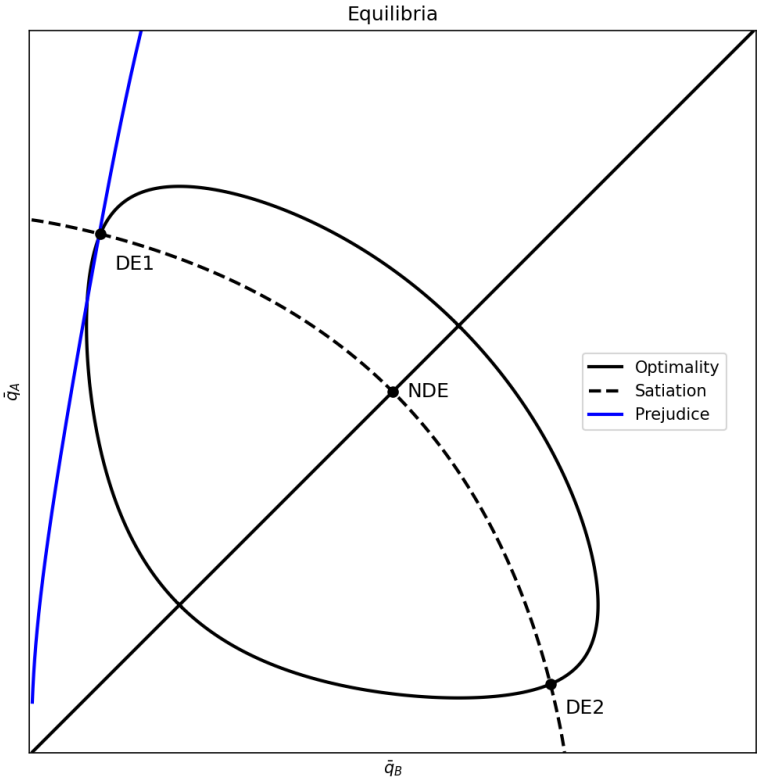

Figure 2: Example of misattribution of equilibrium discrimination under incorrect beliefs of employer type. The locus labelled as *Satiation* plots the combination of application strategies that lead to hiring strategies which jointly satisfy labour demand. The loci labelled *Optimality* plots the combination of application strategies that lead to hiring strategies which jointly satisfy unconstrained profit maximisation. The locus *Prejudice* maps the application strategies that are incentivised by self-confirming beliefs of a naïve and biased employer of relative preference $\hat{m}$.

some $\hat{m} \neq 1$ such that the equilibrium conditions in definition 2 hold.

One can interpret the cursed equilibrium in the following way. Suppose that workers see that the employer only hired A workers, they may then believe that the employer has a taste for A workers and so B workers are less inclined to apply since their believed return to applying is lower than A workers. However, if workers do not correctly estimate that this behaviour is common among all other workers, then they may see the same hiring outcomes again and assume that this is because the employer discards B applicants instead of correctly understanding that this is because the calibre of B applicants is lower on average than A workers because of their application strategy.

Proofs

Lemma 1

An individual i from group g with ability q_i is indifferent to applying to the job if the following function holds,

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

The gradient of the LHS is equal to 1 and so for q_i to be unique given a finite signal threshold $\underline{s}_g$, I need to show that there is a unique fixed point of $y(q : \underline{s}_g) = (1/2)\omega(1 + \text{erf}(\sqrt{\eta/2}(q - \underline{s}_g)))$. The range of $y(q : \underline{s}_g)$ is $(0, w)$ and q has full support so there must be at least one fixed point of the function. The remaining step is to show that the fixed point is unique. To do this, I will show the condition under which $y(q : \underline{s}_g)$ has a gradient in the interval $(0, 1)$ for all $q \in \mathbb{R}$ when assumption 1 holds.

The first order derivative of $y(q : \underline{s}_g)$ can be written as the following,

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

The exponential is always less than one but positive and so, given that $\omega > 0$, the gradient is always positive. Thus to complete the result, a sufficient condition is that $\omega\sqrt{\eta/2\pi} \leq 1$. This is the condition laid out in assumption 1. ■

Lemma 2

The domain of $\chi(\cdot)$ is determined by the domain of the (finite) inverse error function which is $(-1, 1)$. Thus the domain of $\chi(\cdot)$ is $(0, \omega)$ which is the feasible set laid out in lemma 1. Therefore one solution to (5) is any $\bar{q}_A = \bar{q}_B$ in this interval.

Given the domain of $\chi(\cdot)$, the bounds of the range are defined by $\chi(0) = \infty$ and $\chi(\omega) = -\infty$. Thus, since the function is continuous over this interval, it must be a globally decreasing function. The function is convex for $q \in Q \subset (0, \omega)$ if there exists two turning points in the function. As the function is globally decreasing, it suffices to show that there is some $q \in (0, \omega)$ such that $\chi'(q) > 0$. The inverse error function has the greatest slope when its value is equal to 0: $q = \omega/2$ and so I show for a sufficient condition whereby $\chi'(\omega/2) > 0$. The derivative of $\chi(\cdot)$ is the following,

$$\chi'(q) = \frac{(\eta + \hat{\tau}(q))(1 + \hat{x}'(q) - \frac{1}{\omega}\sqrt{\frac{2\pi}{\eta}}e^{\text{erf}(1-\frac{2q}{\omega})^2}) - \hat{\tau}'(q)(q + \hat{x}(q) + \sqrt{\frac{2}{\eta}}\text{erf}(1 - \frac{2q}{\omega})}{(\eta + \hat{\tau}(q))^2}.$$

Thus, the derivative $\chi'(\omega/2)$ can be shown to be the following where $t = \hat{\tau}(\omega/2)$ and $u = \hat{\mu}(\omega/2)$ for convenience,

$$\chi'(\omega/2) = \frac{1 - \frac{1}{\omega}\sqrt{\frac{2\pi}{\eta}}}{t + \eta} + \frac{u't + u'\frac{t^2}{\eta} + t'u - t'\frac{\omega}{2}}{(t + \eta)^2}.$$

From assumption 1, the first component is always non-positive. The second component is positive when the following holds,

$$u't\left(1+\frac{t}{\eta}\right) > t'\left(\frac{\omega}{2}-u\right). \tag{10}$$

To show a condition for when the left hand side is positive, I need to use lemma 2.1.

Lemma 2.1: Dynamics of the mean of the truncated normal distribution. The mean of a normal distribution $N(\mu, 1/\tau)$ truncated from above at α is non-decreasing in α when $\alpha \geq \mu$.

Proof. The described means of the truncated normal distributions are

$$\hat{\mu}(a) = \mu - \frac{\theta(z(a))}{\sqrt{\tau}} \tag{11}$$

where

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

The derivative of $\theta(z(a))$ is the following,

$$\theta'(z(a)) = -\sqrt{\tau}\frac{\psi'(z(a))c(z(a)) + c'(z(a))\psi(z(a))}{c(z(a))^2}.$$

Using the functional forms of $\psi(\cdot)$ and $c(\cdot)$, I can find their derivatives $\psi'(z(a)) = -z(a)\psi(z(a))$ and $c'(z(a)) = \psi(z(a))$. Therefore, it must be that $\theta'(z(a)) = -\sqrt{\tau}(z(a)\theta(z(a)) + \theta(z(a))^2)$. Which is non-positive when $a \geq \mu$.

Differentiating $\hat{\mu}(a)$ with respect to a gets

$$\frac{d\hat{\mu}(a)}{da} = -\frac{\theta'(z(a))\frac{dz(a)}{da}}{\sqrt{\tau}} = -\theta'(z(a))$$

which by the above is non-negative when $z(a) \geq 0$ or when $a \geq \mu$. ■

Thus given lemma 2.1, the LHS of (10) is non-negative when $\bar{q} \geq \mu$. Thus a sufficient condition for $\chi'(\omega/2) > 0$ for some value of ω is that the RHS is negative. To show this, we require lemma 2.2.

Lemma 2.2: Dynamics of the variance of the distribution. The concentration, $\hat{\tau}$ of the normal distribution $N(\mu, 1/\tau)$ truncated from above at α is decreasing in α when $\alpha > \mu$.

Proof. The described concentration is equal to

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

Given the definition of the truncation and distributional variance, it must be that when $z(a) > 0$, $\hat{\tau} > \tau$. The concentration $\hat{\tau}$ has the following derivative,

$$\hat{\tau}'(a) = \frac{\tau\sqrt{\tau}\theta(z(a))}{(1-\theta(z(a))^2 - z(a)\theta(z(a)))^2}\left[1 - (2\theta(z(a)) + z(a))(\theta(z(a)) + z(a))\right].$$

Since the fraction on the RHS of the above is positive, the sufficient condition for $\hat{\tau}'(a)$ to be negative is then that the critical expression $(2\theta(z(a)) + z(a))(\theta(z(a)) + z(a))$ is greater than 1. Since $z(\mu) = 0$, the value of the critical expression is $2\theta(0)^2 = 4/\pi > 1$. The derivative of the critical expression with respect to a is $\sqrt{\tau}\left[(2\theta(z(a)) + z(a))(\theta'(z(a)) + 1) + (\theta(z(a)) + z(a))(2\theta'(z(a)) + 1)\right]$ which is positive if the following holds,

$$\frac{3\theta(z(a)) + 2z(a)}{\theta(z(a))(z(a) + \theta(z(a)))^2} > 1 + \frac{3\theta(z(a)) + 2z(a)}{z(a) + \theta(z(a))}.$$

Since the concentration of the truncated distribution must be positive, then it must be that $1 - \theta(z(a))(z(a) + \theta(z(a))) > 0$ for all a. This can be rearranged to $\theta(z(a)) < 1/(z(a) + \theta(z(a)))$. This means that

$$\frac{3\theta(z(a)) + 2z(a)}{\theta(z(a))(z(a) + \theta(z(a)))^2} > \frac{3\theta(z(a)) + 2z(a)}{z(a) + \theta(z(a))} > 1 + \frac{3\theta(z(a)) + 2z(a)}{z(a) + \theta(z(a))}.$$

Which must hold when $z(a) > 0$. Thus the sufficient condition holds and the lemma is proven. ■

When $\bar{q} \geq \mu$, then the expression (10) holds. Now, it remains to show that $\chi'(\omega/2)$ is greater than zero for some values of parameters.

Suppose that $\omega = \sqrt{2\pi/\eta}$ i.e. assumption 1 binds. Then the first fraction is equal to 0. We have shown then that if $\omega/2 \geq \mu$, the second fraction is positive which satisfies this condition. Since the derivative of $\chi(\cdot)$ is continuous in these parameters by definition, there must exist some values of $\omega \geq 2\mu$ that satisfy assumption 1 and $\chi'(\omega/2) > 0$. For this to be consistent with assumption 1, then it must be that $\mu < \sqrt{\pi/2\eta}$ such that there is some possible values of ω that can satisfy this condition.■

Proposition 1

By lemma 2, the beliefs $\bar{q}_A = \bar{q}_B$ may be simultaneously self-confirming and optimal. If application strategies are identical then (6) becomes the following,

$$\phi = \int_{-\infty}^{\bar{q}} P\left(e \geq \bar{q} - q - \sqrt{\frac{2}{\eta}}\text{erf}^{-1}\left(\frac{2\bar{q}}{\omega} - 1\right)|q\right) f_q(q)dq.$$

The RHS is monotonically increasing in $\bar{q}$ from 0 when $\bar{q} = 0$ to $F_q(\omega)$ when $\bar{q} = \omega$. Thus, if $\phi \leq F_q(\omega)$, then there will exist some pair of beliefs $(\bar{q}_A, \bar{q}_B)$ which simultaneously satisfy (6), (5) and $\bar{q}_A = \bar{q}_B$ to constitute a non-discriminatory equilibrium. ■

Proposition 2

In lemma 2, I show that when $2\mu \leq \omega \leq \sqrt{2\pi/\eta}$, there is some set $\tilde{Q} \subset (0, \omega)$ such that if $q \in \tilde{Q}$, then there also exists some $q', q'' \in \tilde{Q}$ such that $\chi(q) = \chi(q') = \chi(q'')$. Since $\chi(q)$ is continuous on the domain $(0, \omega)$ and globally decreasing, it must be that $\tilde{Q}$ is a continuous set meaning $\tilde{Q} = (q_m, q^m)$.

Partition $\tilde{Q}$ into three subsets: $\tilde{Q}_0 = (q_m, q_1)$, $\tilde{Q}_1 = [q_1, q_2]$ and $\tilde{Q}_2 = (q_2, q^m)$ such that if $q \in \tilde{Q}_0$, then by lemma 2, it must be that $q' \in \tilde{Q}_1$ and $q'' \in \tilde{Q}_2$. Moreover, it must be that $\chi(q_m) = \chi(q_2)$ and $\chi(q^m) = \chi(q_1)$. From proposition 1, if $\phi \leq F_q(\omega)$, then there will be a non-discriminatory equilibrium defined by $\bar{q}$. Since (6) is increasing in $\bar{q}$, that means that $\bar{q}$ is increasing in ϕ from 0 to $F_q(\omega)$. As such, if $\tilde{Q}$ exists, there is some set $\Phi \subseteq (0, F_q(\omega))$ such that $\bar{q} \in \tilde{Q}$. Hence by the same logic, Φ can be partitioned into 3 subsets such that if $\phi \in \Phi_i$ then $\bar{q} \in \tilde{Q}_i$ for $i \in \{0, 1, 2\}$.

According to the above, then we can define the condition in (7) on the $(\bar{q}_A, \bar{q}_B)$ plane by the following. The 45 degree line constitutes pairs of $(\bar{q}_A, \bar{q}_B)$ from $(0, \omega)$ that satisfy (7). Moreover, when $2\mu \leq \omega \leq \sqrt{2\pi/\eta}$ two loci of points can be defined which constitute pairs of self-confirming beliefs that satisfy (7). These loci are mirror images over the 45 degree line due to the symmetric nature of (7). Each locus connects the points $(\bar{q}_1, \bar{q}_1)$ and $(\bar{q}_2, \bar{q}_2)$ via the pairs of $\bar{q}_A$ and $\bar{q}_B$ that satisfy (7) such that $\bar{q}_A \neq \bar{q}_B$. Intuitively, this draws a continuous oval passing through the points $(\bar{q}_1, \bar{q}_1)$, (q_m, q_2), (q_1, q^m), $(\bar{q}_2, \bar{q}_2)$, (q^m, q_1) and (q_2, q_m).

On the $(\bar{q}_A, \bar{q}_B)$ plane, (6) is decreasing. By proposition 2 when $\phi \leq F_q(\omega)$ it passes through the 45 degree line. Thus if (6) is defined at $\bar{q}_g = 0$ for both $g \in \{A, B\}$ and it passes through the 45 degree line at (q, q) where $q \in \tilde{Q}_1$, then it must pass through (5) twice more constituting the non-discriminatory equilibria. By lemma 1, then for this to exist, it must be that $\phi/n_g \leq F_q(\omega)$ for each $g \in \{A, B\}$. Thus if μ simultaneously satisfies the following conditions then non-discriminatory equilibria exist:

$$0 \leq 2\mu \leq \omega \leq \sqrt{\frac{2\pi}{\eta}}, \tag{12}$$

and

$$\frac{\phi}{\min\{n_A, n_b\}} \leq F_q(\omega). \tag{13}$$

I first find a condition such that there is some positive wage ω that satisfies (13). By assumption 1, this sets the following lower bound for μ such that, without loss of generality $n_A \leq n_B$,

$$\mu > \sqrt{\frac{2}{\tau}} \text{erf}^{-1}\left(1 - \frac{2\phi}{n_A}\right). \tag{14}$$

Moreover, it must be that the two conditions are consistent: that there is a set of ω that satisfies (12) and (13). For there to exist some ω to satisfy (12), it must be that $\mu < \sqrt{\pi/2\eta}$. Moreover, for (13), it must be that

$$\mu + \sqrt{\frac{2}{\tau}} \text{erf}^{-1}\left(\frac{2\phi}{n_A} - 1\right) < \sqrt{\frac{2\pi}{\eta}},$$

which implies the following

$$\phi < \frac{n_A}{2}\left(1 + \text{erf}\left(\frac{1}{2}\sqrt{\frac{\pi\tau}{\eta}}\right)\right). \tag{15}$$

Thus it must be that (15) and (14) are compatible for there to exist some value of ω such that there are multiple equilibria. It can be shown that for these conditions to be consistent, it must be that $\mu < \sqrt{\pi/2\eta}$ which is (12) and $\phi > (n_A/2)(1-\text{erf}(1/2\sqrt{\pi\tau/\eta}))$ which provides a lower limit to ϕ. Thus when the following conditions hold, it must be that there is some permissible value of $\omega > 0$ such that the non-discriminatory equilibria are possible:

$$\phi \in \Phi \equiv \left(\frac{n_A}{2}\left(1 - \text{erf}\left(\frac{1}{2}\sqrt{\frac{\pi\tau}{\eta}}\right)\right), \frac{n_A}{2}\left(1 + \text{erf}\left(\frac{1}{2}\sqrt{\frac{\pi\tau}{\eta}}\right)\right)\right), \tag{16}$$

and

$$\mu \in \mathbb{M} \equiv \left(\sqrt{\frac{2}{\tau}}\text{erf}^{-1}\left(1 - \frac{2\phi}{n_A}\right), \sqrt{\frac{\pi}{2\eta}}\right). \tag{17}$$

■

Lemma 3

Given lemma 1, the value $\bar{q}_g$ is monotonic in $\underline{s}_g$. Given some belief of the signal threshold applicable, as $\hat{m}$ increases, it must be that in order to satisfy the labour demand, $\underline{s}_A/\underline{s}_B$ is decreasing. Since $\bar{q}_A/\bar{q}_B$ is decreasing in the ratio of signal thresholds, this means that it is increasing in the belief $\hat{m}$. When $\hat{m} = 1$, the workers believe the employer is fair and applies signal thresholds $\underline{s}_A = \underline{s}_B$, thus workers do not believe that they will be discriminated against and therefore adopt a pooling strategy. ■

Proposition 3

When $\hat{m} = 1$ and $\hat{\psi} = 0$, the workers believe the employer will always set equal signal thresholds and so employ the application strategy defined by the intersection of locus $\bar{q}_A = \bar{q}_B = \bar{q}$ and (7). This is also the definition of the profit maximising non-discriminatory equilibrium defined in proposition 1. ■

Proposition 4

Definition 2 outlines an equilibrium in which workers have stable but inaccurate beliefs about the employer type in an equilibrium where their beliefs of hiring strategies are self-confirmed by a constrained profit maximising employer. Of course, nested in this equilibrium definition is the Perfect BNE shown in the previous section. I am now interested in the case of $\hat{\psi} = 0$ and $\hat{m} > 1$.

In this case, the workers believe that they will face signal thresholds that simultaneously satisfy $\hat{m} = \underline{s}_A/\underline{s}_B$ and (6). Denote the application strategy which is a best response to the belief $\hat{m}$ as $(\bar{q}_A^{\hat{m}}, \bar{q}_B^{\hat{m}})$. Accordingly, it must be that the following holds,

$$\hat{m} = \frac{\bar{q}_A^{\hat{m}} - \sqrt{\frac{2}{\eta}}\text{erf}\left(\frac{2\bar{q}_A^{\hat{m}}}{\omega} - 1\right)}{\bar{q}_B^{\hat{m}} - \sqrt{\frac{2}{\eta}}\text{erf}\left(\frac{2\bar{q}_B^{\hat{m}}}{\omega} - 1\right)}. \tag{18}$$

Under assumption 1, this is a monontically increasing function on the $(\bar{q}_A, \bar{q}_B)$ plane for a finite $\hat{m}$. When $\hat{m} = 1$, this function is only satisfied by $\bar{q}_A = \bar{q}_B$. In other words, when the workers do not believe in any taste based discrimination, the only signal thresholds that would satisfy these beliefs are equal. By lemma 3, the distance between (18) and the 45 degree line increases as $\hat{m}$ increases. Moreover, by proposition 2, when $\phi \in \Phi$ and $\mu \in \mathbb{M}$, then (6) is monotonically decreasing on this plane between the two axes. Thus the intersection point of (18) and (6) defines the application strategy that workers employ and therefore as $\hat{m}$ increases, $\bar{q}_A$ decreases and $\bar{q}_B$ increases. ■

Proposition 5

In the case outlined, workers will believe that they will face signal thresholds which satisfy $\hat{m}\underline{\hat{s}}_A = \underline{\hat{s}}_B$. Accordingly, this lines up with the conditions described in Proposition 4 and thus the proof is identical to this. ■

Instructions

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 study 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 study 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. If you disagree, you may withdraw now by raising your hand.

[ ] I agree to participate in this study and understand how the data I provide will be used.

Page 2 — Overview of the study

In this study, you will play 12 rounds of a hiring game.

In each round, you will make 3 different kinds of decisions, which 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 10, 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.**

You can follow along with the instructions on screen. A print-out of the instructions is distributed to you and a short summary is on the back of the handout.

Page 3 — Your Society and Your Colour (Instructions Page 1/4)

Your society For the study, 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).

Page 4 — 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.

The Productivity Urn:

- Imagine an urn called the **Productivity Urn** which contains **1,000 balls**, each with a **whole** number written on it (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.

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

Your productivity value for round 1 *You will now see your productivity value for round 1. In each of the following rounds, you will draw a new productivity value.*

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 **not to apply** for a job, your bonus 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.

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

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

Page 5 — 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.

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.

- 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.

An example calibrated to your productivity value is displayed below.

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 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.

Key points:

- For each applicant, the employer sees only their **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.

Page 7 — The Employer (Instructions Page 4/4)

The employer uses a computerised 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 modelling**:

- 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:

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

On the next page is a short quiz to check your understanding of the instructions on this page. Afterwards you will see a break-down of the decisions you have to make in each round and then the rounds will begin.

Page 9 - Your Decisions

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

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 on your screen, for instance, 1 green worker and 2 purple workers were hired in round 4.

Below, you will read about each decision that you must make in each round and the corresponding payment rules.

Decision 1: 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.

Decision 2: You will guess the maximum productivity value among the participants in your soceity who chose to apply in the current round, separately for

- Green participants (excluding you if you apply).
- 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 guess:

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

Decision 3: 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 earning.
- 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.
  - Say that you allocate Q tokens to robot A being used. Let q be Q divided by 100 e.g. if you allocate 60 tokens to robot A, then $q = 0.6$.
  - If robot A is being used, you will earn $100[1 - (1 - q)^2]$ tokens. In the above example, if robot A is being used you will earn 84 tokens.
  - If robot B is being used, you will earn $100(1 - q^2)$ tokens. In the above example, if robot B is being used, you will earn 64 tokens.
  - This formula was proven to align your interests with truthful reporting in the paper by Brier, G.W in 1950 titled "Verification of forecasts expressed in terms of probability".

You must wait for all individuals in your society to complete the round before moving to the next one.

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.

SUMMARY

- You have been assigned to a society of 16 people in this room.
- You have been randomly allocated a colour.
- Society and colour assignment stay for the entirety of the study.

In each round:

- You draw a productivity value from an urn (average value = 50, 25% draws are below -5 and 25% of draws are above 105.).
- **DECISION 1:** You choose whether to apply for a job.
- If you do not apply, your decision 1 bonus = your productivity value.
- If you apply:
  - You go into an applicant pool with others in your society who also apply.
  - You draw a ball from the noise urn (average noise = 0) which generates your signal

    $$\text{your signal} = \text{productivity value} + \text{draw from noise urn}$$

  - A robot chooses 3 participants from the applicant pool to hire.
  - The robot can be robot A (systematic adjuster) or robot B (advanced statistical modeller).
  - The robot employed remains the same for your society in all rounds.
  - The robot observes applicants' colour and signal.
  - If you are hired, you earn 90.
  - If you are not hired, you earn 0.
- **DECISION 2:** You guess the **maximum** productivity value of **green participants in the applicant pool** and of **purple participants in the applicant pool**.
- **DECISION 3:** You observe the colours of those hired and **guess which robot** was used.

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.

Simulations

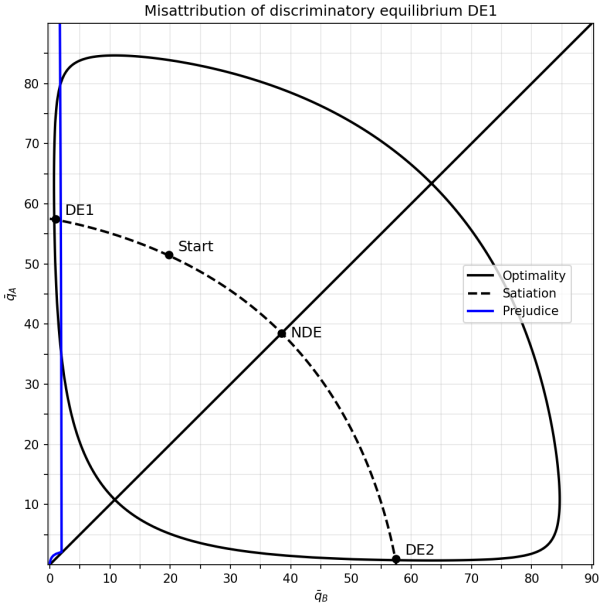

Figure 3: Simulated equilibria and starting prior belief of robot A. Non-discriminatory equilibrium is labelled 'NDE' and discriminatory equilibria are labelled 'DE1' and 'DE2'. The prejudice locus is mapped on the y-axis in this domain identifying the set of application strategies that are optimal when employer is believed to taste-discriminate. 'DE1' is the approximate non-discriminatory equilibrium which can be misattributed to prejudice as well as with statistical discrimination. The start point is roughly half-way along the satiation curve between the discriminatory equilibrium congruent with both statistical discrimination and taste-based discrimination and the non-discriminatory equilibrium.

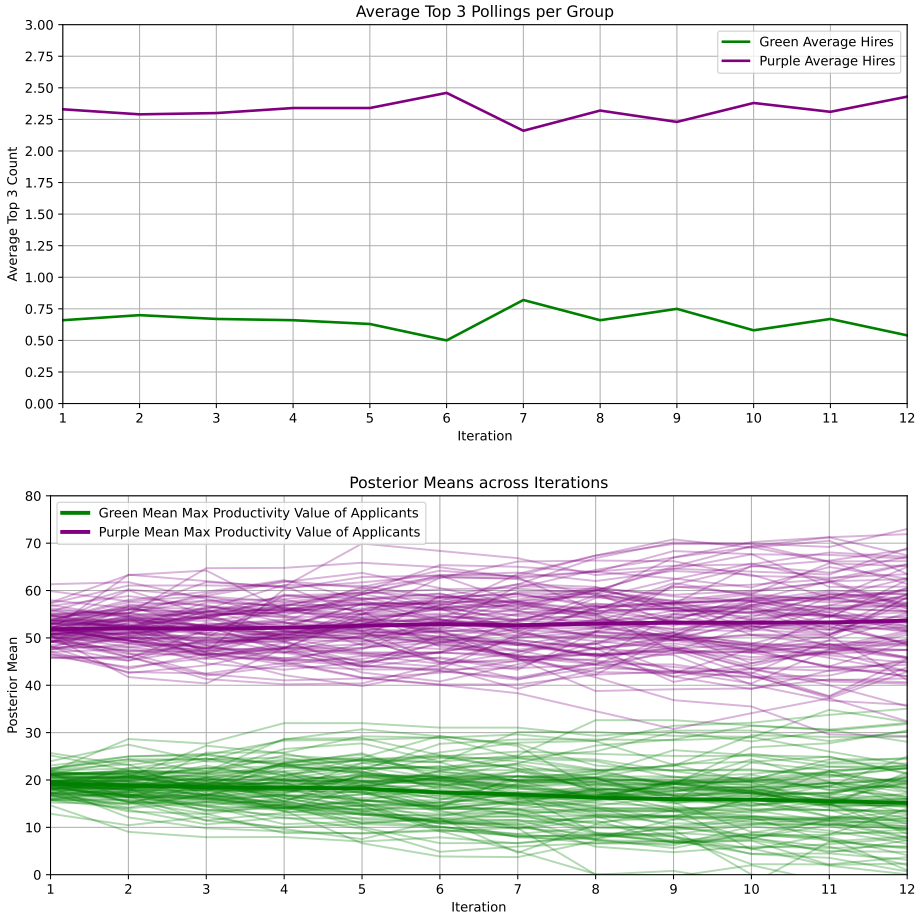

Figure 4: Simulated experiment outcomes with starting point outlined in Figure 3 and prior standard deviation 15. Simulated application strategy is congruent with DE1 where greeb workers apply if and only if their productivity value is less than or equal to 1 and purple workers apply if and only if their 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.

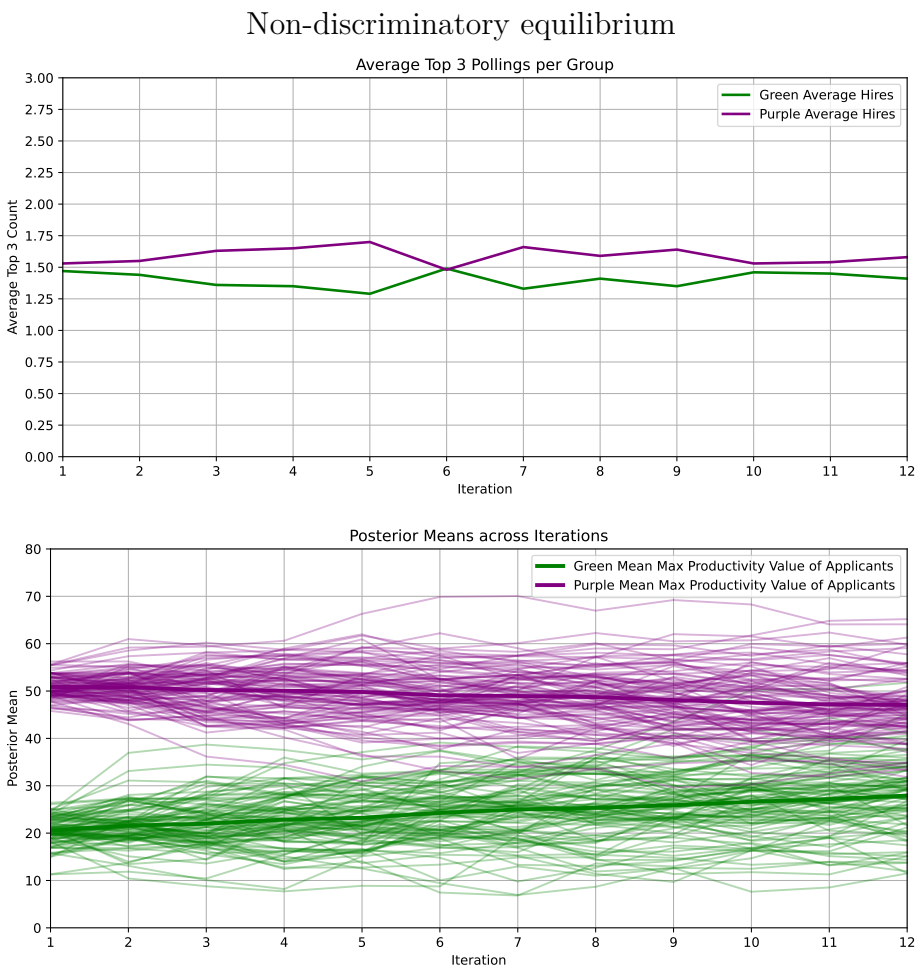

Figure 5: Simulated experiment outcomes with starting point outlined in Figure 3 and prior standard deviation 15. Simulated application strategy is congruent with NE where both groups of workers apply if and only if their productivity value is less than or equal to 38. Upper graph shows the simulated 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.

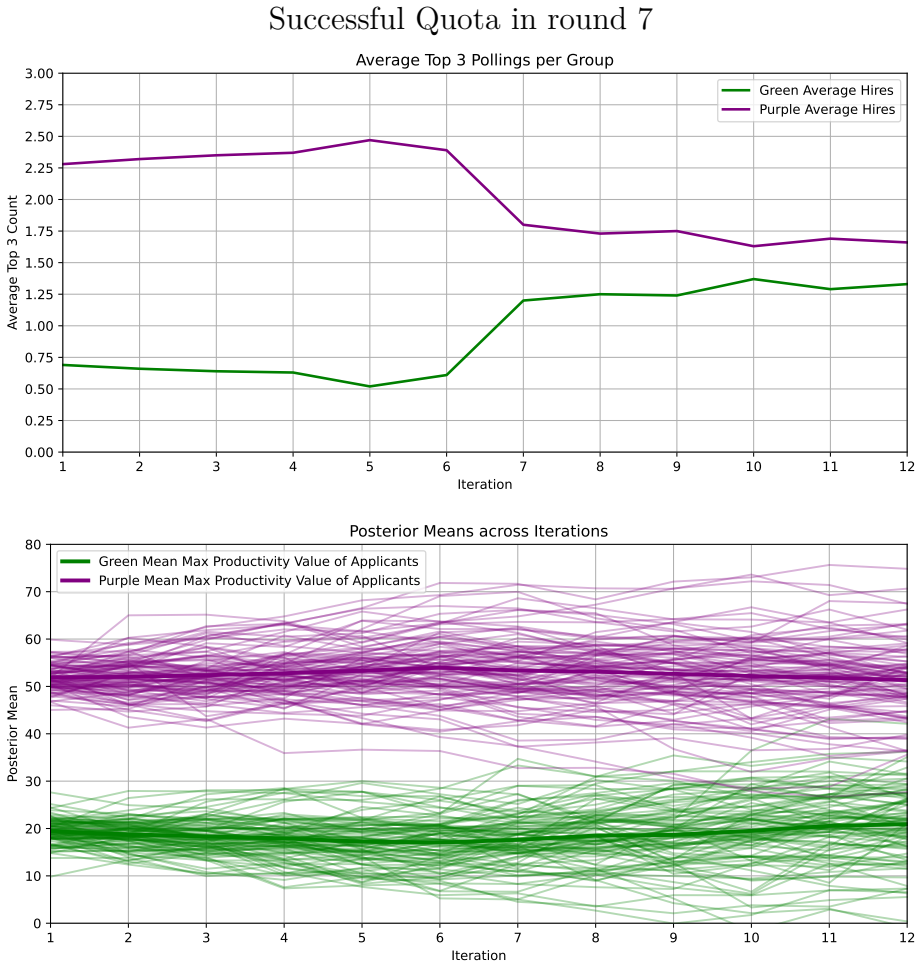

Figure 6: Simulated experiment outcomes with starting point outlined in Figure 3 and prior standard deviation 15. Simulated application strategy is congruent with DE1 in the first 6 rounds and NE in the last 6 rounds. This is akin to a quota being enforced in round 7 and thereafter strategies converge 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.

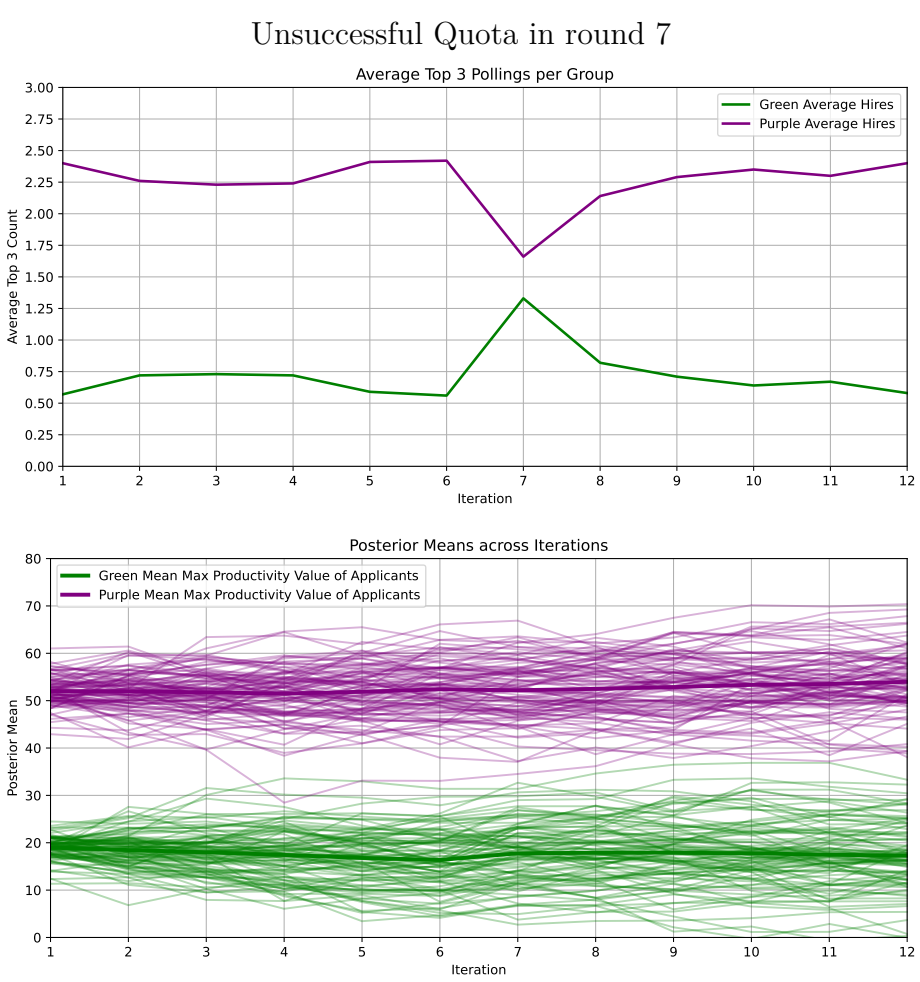

Figure 7: Simulated experiment outcomes with starting point outlined in Figure 3 and prior standard deviation 15. Simulated application strategy is congruent with DE1 except in round 7 where a quota is implemented and NE is enforced. Upper graph shows the simulated average rate of hiring green and purple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