Explainable AI and Human Decision Making: Preferences, Beliefs, and Biases

Peter Bergman   Tushar Kundu   Kadeem Noray

UBC VSE Empirical Lunch

December 2, 2024

AI is Gatekeeper for Economic Mobility

AI serves as an agent for economic decision making

- Resume screening and hiring
- Loans and credit decisions
- Healthcare access and coverage
- Housing applications

LLMs have accelerated adoption

- Easy to implement: pre-trained models
- Capable of mimicking human behavior (e.g. Horten et al. 2024)

But raises concerns

- Complex, black-box decision making
- Hard to explain decisions
- Difficult to parse sources of bias

Key Questions:

- Can we use models of human behavior to explain AI behavior?
- How well do GenAI and humans assess candidate quality?
- How do AI vs. human evaluations differ?
  - Preferences over candidates
  - Beliefs about quality
  - Types of biases

Challenging to answer

The selective labels problem

- Only observe outcomes for accepted candidates

Hard to separate multiple sources of bias

- Taste-based discrimination
- Biased beliefs
- Statistical discrimination
- Decision-maker heterogeneity (cf. Kline, Rose, and Walters 2021)

Our Approach

- Partner with interviewing.io, a platform where users conduct technical interviews
- Ask human recruiters and AI to evaluate resumes of platform users
- Compare AI vs human decisions
- Model resume evaluation decision making to identify discrimination sources
  - Separate beliefs from preferences

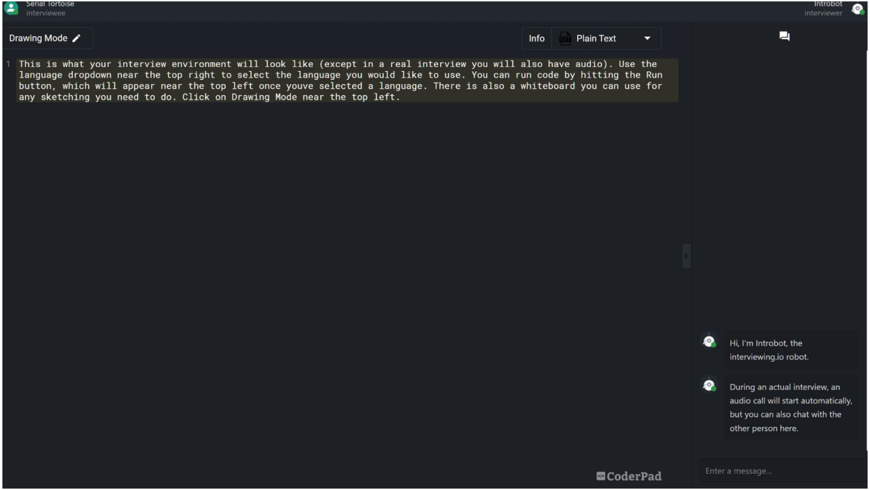
Novel dataset and setting

- Observe true candidate quality
- Compare AI vs. human decisions

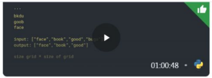
Key innovations

- No selective labels problem
- Model different forms of discrimination
- Quantify decision-maker heterogeneity
- Assess relative AI performance

Interviewing.io



Interviewing.io



Aerodynamic Raven
interviewed by Clandestine Hamburger

02/05/22 at 3:00pm • Algorithms/Data Structures

PRO • Interviewer from Facebook

Feedback about Aerodynamic Raven (the interviewee)

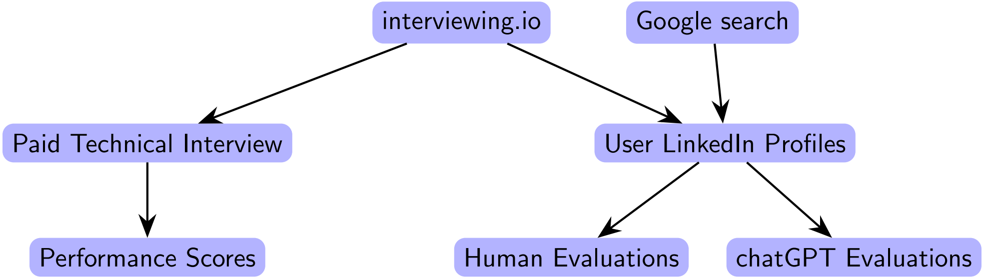
Advance this person to the next round?	yes
How were their technical skills?	★★★★
How was their problem solving ability?	★★★★
What about their communication ability?	★★★★

> *Great communication, problem analysis, problem solving, coding and testing. I really liked how you explained your test plan, and organizing code into helper functions.*
>
> *Asking for clarifying questions such as constraints and edge cases will help with solving problems.*
>
> *Most interviewers will ask you 2 problems so you have about 20m to solve each of them.*
>
> *Suggest practicing a few problems for every kind of DS & algo: string, array, 2-pointer (palindrome Q), stack, linked list, hash, DFS/BFS, tree/trie/BST, queue, heap.*
>
> *Good luck!*

Feedback about Clandestine Hamburger (the interviewer)

Would you want to work with this person?	yes
How excited would you be to work with them?	★★★★
How good were the questions?	★★★★
How helpful was your interviewer in guiding you to the solution(s)?	★★★★

Data Collection Overview



- Two key data sources:
  - Actual interview performance (ground truth)
  - Resume evaluations by humans and AI

Step 1: Platform Data Collection

Interview Performance Metrics

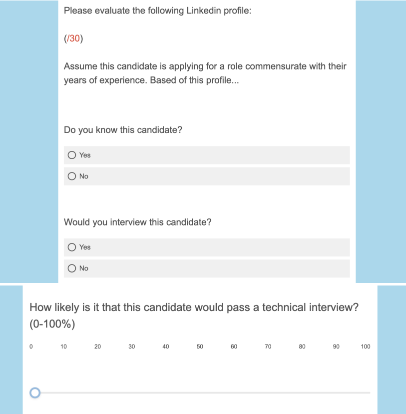
- Technical abilities
  - Coding skills (1-4 scale)
  - Problem-solving (1-4 scale)
- Soft skills
  - Communication (1-4 scale)
- Overall assessment
  - Would hire (Yes/No)

Candidate Information

- LinkedIn profiles
  - Education
  - Work experience
  - Certifications
- Demographics
  - Gender (inferred from names/photos)
  - Race (inferred from names/photos)

Step 2: Human Recruiter Evaluation

- Surveyed 78 professional technical recruiters
    - Firms include Amazon, Meta, Microsoft, Stripe, etc.
    - Paid $2.50 per evaluation
    - Incentivized on accuracy ($1.50 if within 10 pp of true pass rate)
- Each recruiter evaluates 30 random profiles
- Two key questions:
    - "Would you interview this candidate? (yes/no)"
    - "How likely is it that this candidate would pass a technical interview on a scale of 0-100%?"



Step 3: ChatGPT Evaluation

Evaluation details:

- Same candidate pool as human recruiters
- Standardized prompt:
  - LinkedIn profile information
  - Identical questions as human recruiters
  - Controlled response format
- Model: gpt-4o
- Perfectly reproducible (temperature $= 0$)

Key features:

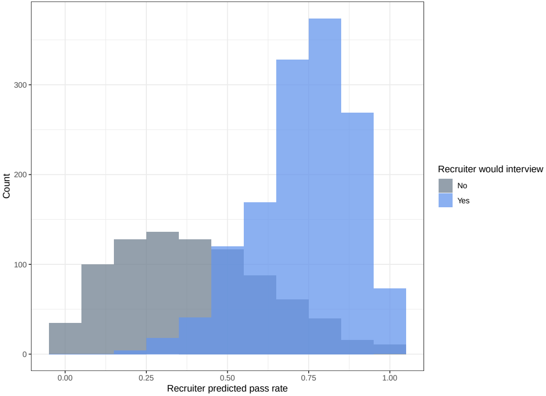
- No access to photos
- Real name included
- Input string includes LinkedIn experience, education, and certification history

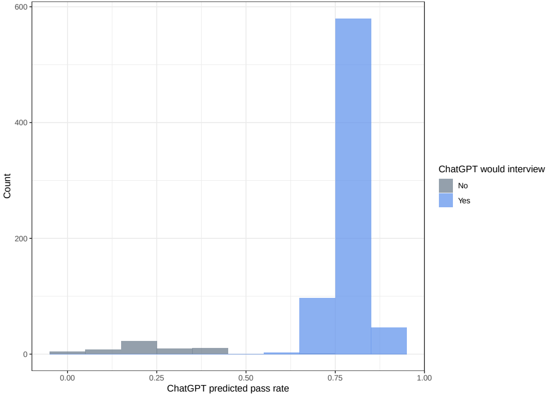
Sample Description

Human Recruiter Evaluations			
Race	Male	Female	Total
white	736	125	861
Black	35	27	62
East Asian	287	109	396
Hispanic	98	11	109
South Asian	550	120	670
Total	1706	392	2098

ChatGPT Predictions			
Race	Male	Female	Total
white	275	35	310
Black	12	5	17
East Asian	130	35	165
Hispanic	21	1	22
South Asian	200	50	250
Total	638	126	764

Interview Decisions and Pass Probabilities





Protected Class Analysis: Interview Recommendations

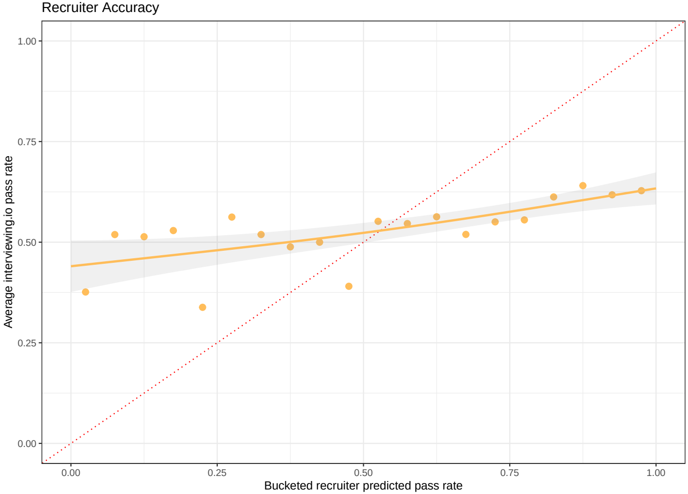
- ChatGPT interviews much higher share of candidates
- Female candidates interviewed less by *humans*
- URM (Black and Hispanic) interviewed more by *humans*

Would Interview $\sim$ Gender $\times$ URM $\times$ Source

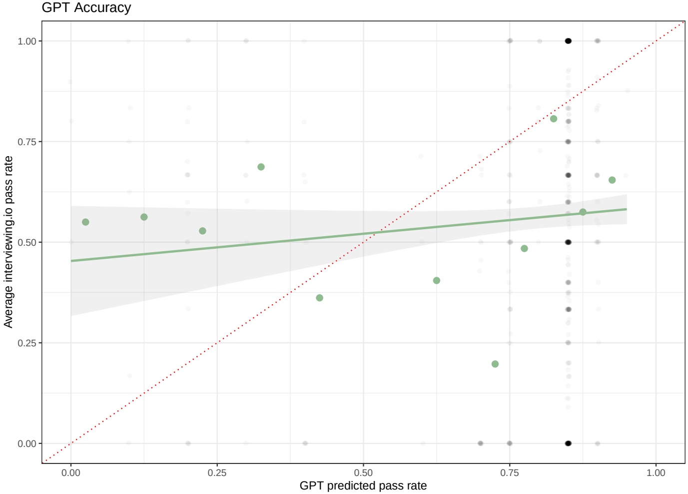
Variable	**Eval. Source**	**Estimate**	**SE**	**P-value**
Overall Marginal Effects				
Recruiter - ChatGPT	Combined	-0.30***	0.015	0.00
URM - non-URM	Combined	0.06	0.036	0.12
Female - Male	Combined	-0.06*	0.025	0.02
ME of URM by Source				
URM - non-URM	ChatGPT	0.00	0.044	0.93
URM - non-URM	Human Recruiter	0.08	0.047	0.09
ME of Gender by Source				
Female - Male	ChatGPT	-0.02	0.028	0.39
Female - Male	Human Recruiter	-0.07*	0.032	0.03

Notes: SE clustered by interviewee. Sig. levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

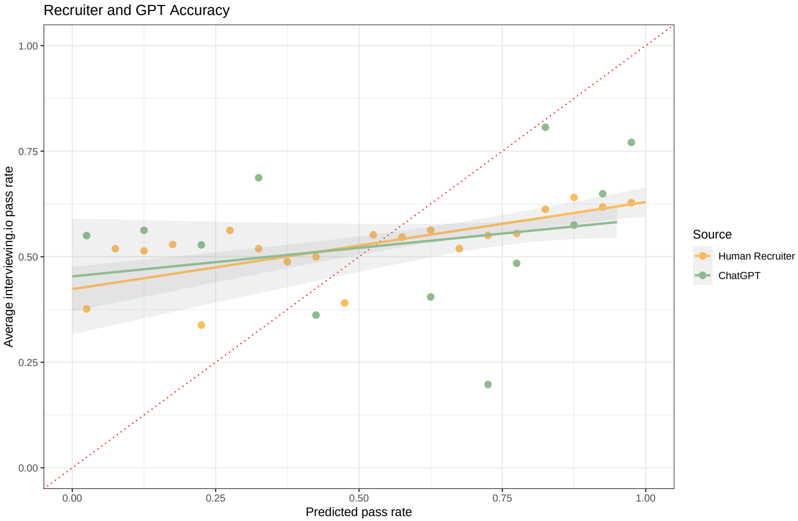
Recruiters are not accurate



ChatGPT isn't accurate either



Recruiters and ChatGPT are similarly not accurate



Model Overview

Goal: Reduced form analysis doesn't identify different types of biases

Three key components:

- True latent quality (unobserved)
- Objective quality measures (interview platform)
- Recruiter/LLM decisions

Key features:

- Separately identifies taste based v. statistical discrimination
- Allows for biased beliefs
- Allows for heterogeneity across recruiters
- Can simulate different policies: blinding candidate characteristics, impacts of eliminating different forms of bias on decision making

True Quality

Latent measure of quality:

$$q_i = \boldsymbol{\delta}' \mathbf{X}_i + \nu_i, \quad \nu_i \sim \mathcal{N}(0, \sigma_\nu)$$

- $\mathbf{X}_i$: Observable resume characteristics (education, experience, etc.)
- δ: True relationship between characteristics and quality
- ν_i: Unobserved component (e.g. soft skills, etc.)

Quality Measures

Technical Interview Performance:

$$M_{ik} = l \quad \text{if} \quad c_{k,l-1} < \phi_k q_i + \epsilon_{ik} \leq c_{k,l}$$

- M_{ik}: Ordinal score on measure k (e.g., coding ability, problem-solving, communication)
- ϕ_k: How well measure k captures true quality
- $c_{k,l}$: Thresholds defining score levels

Final Hiring Decision:

$$h_i = \mathbb{1}\{\phi_{\text{hire}} q_i + \xi_i > c_{\text{hire}}\}$$

- Binary outcome: hire/no hire

Recruiter Beliefs and Preferences

Beliefs about quality:

$$q_i \sim \mathcal{N}(\boldsymbol{\alpha}'\mathbf{X}_i, \sigma_\nu)$$

- $\boldsymbol{\alpha}$: Recruiter's beliefs about how characteristics predict quality
- Can differ from true relationship ($\boldsymbol{\delta}$) $\rightarrow$ Biased beliefs

Utility from interviewing:

$$U_{ij} = [\boldsymbol{\beta}_j'\mathbf{X}_i] + \gamma_j[\boldsymbol{\alpha}_j'\mathbf{X}i] + \varepsilon_{ij}$$

- $\boldsymbol{\beta}_j$: Direct preferences over non-quality characteristics (taste-based discrimination)
- γ_j: Weight on expected quality
- Heterogeneity across recruiters (j subscript)

Interview Decisions and Predictions

Interview Decision:

$$\text{Interview}_{ij} = \mathbb{1}\{U_{ij} > \tau_j\}$$

- Interview if utility exceeds recruiter-specific threshold
- Combines preferences and beliefs

Pass Prediction:

$$p_{ij} = \Phi(\lambda([\boldsymbol{\alpha}_j' \mathbf{X} i] - \mu_j) + \eta_{ij})$$

- p_{ij}: Probability recruiter thinks candidate will pass
- Based *only* on beliefs about quality ($\boldsymbol{\alpha}_j$)
- μ_j: Recruiter specific perception of required quality

Sources of Discrimination

Model captures three distinct channels:

Taste-Based

- Direct preferences (β)
- Unrelated to productivity

Statistical

- Using characteristics to predict quality
- Based on correct beliefs ($\boldsymbol{\alpha} = \boldsymbol{\delta}$)

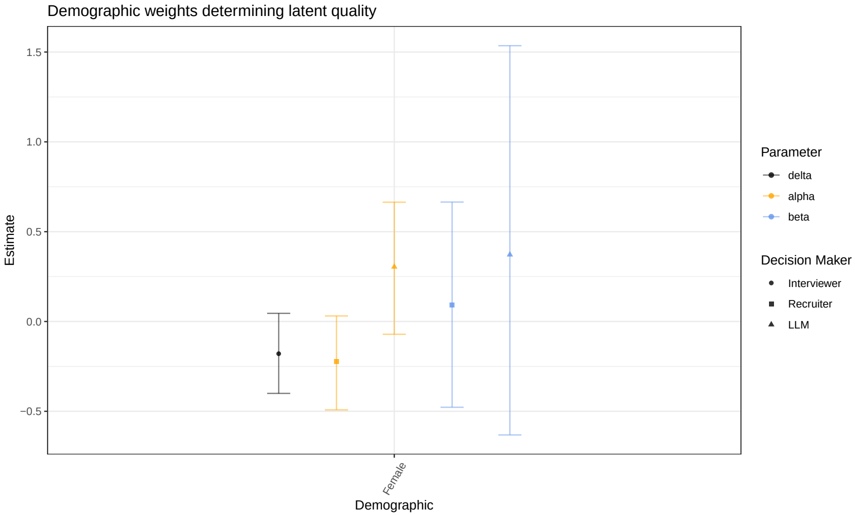
Biased Beliefs

- Incorrect quality predictions
- When $\boldsymbol{\alpha} \neq \boldsymbol{\delta}$

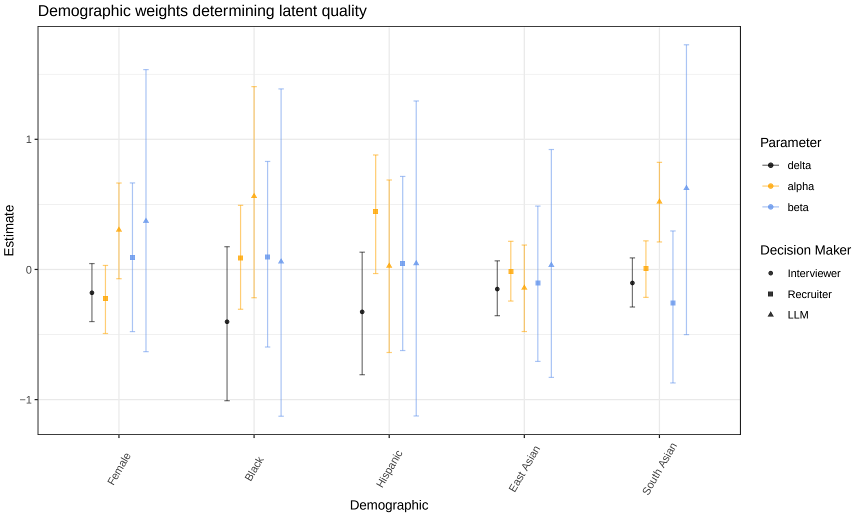
Key Insights:

- Model allows us to separately identify these sources by comparing beliefs to true relationships
- e.g. reduced form might identify $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = \gamma_j \boldsymbol{\alpha}_j + \boldsymbol{\beta}_j$, which is a function of preferences and beliefs
- Hinges on our ability to (1) measure latent quality and (2) measure beliefs

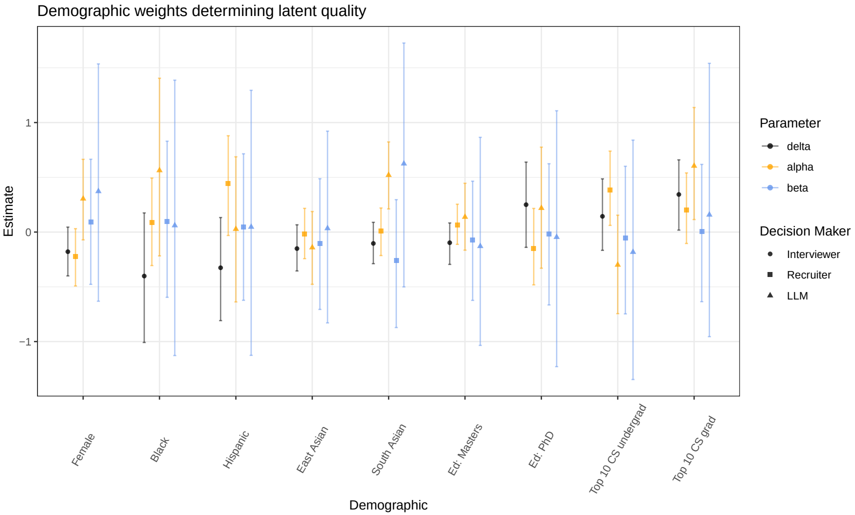
Model Results



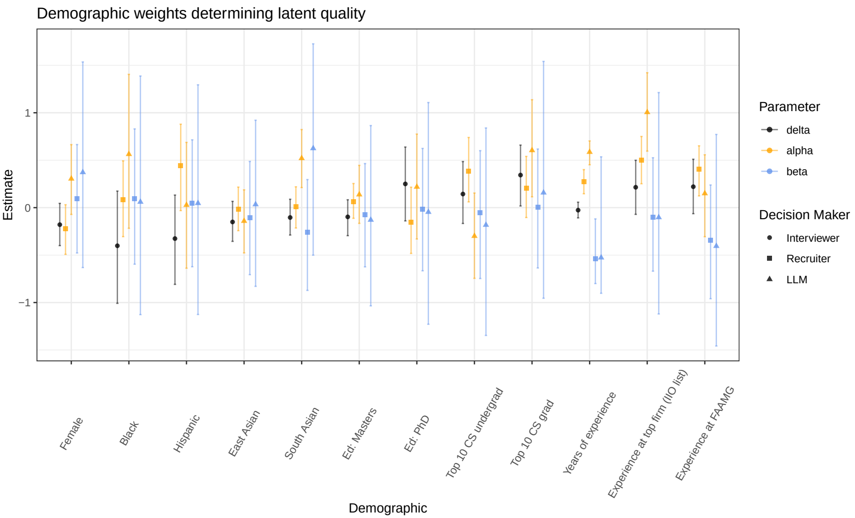
Model Results



Model Results



Model Results



Next Steps: Model Incorporating Provision of Algorithmic Score

Recruiter's receive quality signal from an algorithmic score:

$$s_i = \boldsymbol{\delta}' \mathbf{X}_i$$

where:

- s_i is the posterior estimate, $\boldsymbol{\delta}$, from predicting q_i using observable resume characteristics $\mathbf{X}_i$

Given our prior distributions, the joint distribution of (s_i, q_i) is bivariate normal:

$$(s_i, q_i) \sim \mathcal{N}\left(\begin{pmatrix} \delta' \mathbf{X}_i \\ \delta' \mathbf{X}_i \end{pmatrix}, \begin{pmatrix} \mathbf{X}_i' \Sigma_\delta \mathbf{X}_i & \mathbf{X}_i' \Sigma_\delta \mathbf{X}_i \\ \mathbf{X}_i' \Sigma_\delta \mathbf{X}_i & \mathbf{X}_i' \Sigma_\delta \mathbf{X}_i + \sigma_\nu^2 \end{pmatrix} \right).$$

How recruiters update their beliefs

Conditional on s_i, we have:

$$q_i \mid s_i \sim \mathcal{N}(s_i,\ \sigma_\nu^2).$$

$q_i \mid s_i \sim \mathcal{N}(s_i, \sigma_\nu^2)$ can be rewritten as:

$$q_i = s_i + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\nu^2).$$

If recruiters are Bayesian, they will update their beliefs:

$$\mathbb{E}[q_i|\mathbf{X}_i, s_i] = \omega s_i + (1-\omega)\boldsymbol{\alpha}_j'\mathbf{X}_i$$

where:

- $\omega = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\nu^2}$ is the weight placed on the signal of quality

What we can do with this extension of the model

Run a (new) experiment providing algorithmic score to answer:

- Collect data on recruiter's prior means and variances over candidates
- Randomize provision of "algorithmic score" (posterior $\delta' \mathbf{X}_i$)
- Do recruiters update their beliefs in a Bayesian way?
- If not, how do recruiters deviate from Bayesian updating?
- How does algorithmic score provision affect biases?
- Use the model to optimally combine information from the algorithm and recruiters
- Compare optimal decision making to recruiters' actual decision making

Summary

Key Findings

- Distinct bias patterns
  - Human recruiters favor URM candidates
  - ChatGPT favors South Asian candidates
- Neither group accurate in predicting performance
- Sources of bias differ
  - Most of the disparate treatment is due to productivity beliefs (alpha)
  - Work experience: Positive statistical discrimination, negative taste based

Next Steps

- Expand sample
  - Focus on underrepresented groups
- Test alternative LLM prompts
  - Fixed pass rate constraint
  - Ranking task
- Resume audit study
- Conduct experiment with algorithmic score provision
- Re-estimate the model

Thank you!

Please reach out to tk2859@columbia.edu with any questions, comments, or suggestions.

We thank our technology partner, **interviewing.io**, and data scientists **Leoson Hoay, Joseph Herrera, and Nitya Raviprakash** of Learning Collider.