Fairness in job recommendations: Estimating, explaining, and reducing gender gaps

Guillaume Bied\textsuperscript{1 2}, Christophe Gaillac\textsuperscript{3}, Morgane Hoffmann\textsuperscript{1}, Philippe Caillou\textsuperscript{2}, Bruno Crépon\textsuperscript{1}, Solal Nathan\textsuperscript{2}, Michèle Sebag\textsuperscript{2}

\textsuperscript{1}Centre de Recherche en Économie et Statistique (CREST), France
\textsuperscript{2}Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), Université Paris-Saclay, France
\textsuperscript{3}Oxford University, United Kingdom

AEQUITAS Workshop at ECAI, September 26th 2023
Outline

Introduction

Related work

Context: data and algorithm

Methodology

Results
- Recommendation performance
- Gender gaps in recommendations

Adversarial debiasing

Conclusion and perspectives
Motivations

- Recommender systems (RS) help users find relevant items in large datasets, leveraging past interactions

- Job recommendation is a key application domain of AI for Good
  - Role of imperfect information in unemployment Belot et al., 2019
  - Highly consequential: jobs determine livelihoods and social positions

- But algorithms trained on real-world data also learn job seekers’ and recruiters’ biases
  - AI in HR: high-risk according to EU AI Act
This work

- Audit of a job RS wrt gender biases
  - Context: partnership with the French Public Employment Service
  - Hybrid RS leveraging rich contextual data on job ads and job seekers
  - Trained on hires

- Goals:
  - Discuss relevant gender gap measures for job recommendation
  - Assess gender gaps in terms of:
    - Performance (recall)
    - Recommended job characteristics: wage, contract, working hours . . .
  - Assess whether the algorithm reproduces / increases disparities present in hiring and application behavior
  - Assess a gender-blind (adversarial) recommender system
    - Cost of neutrality in terms of recall ?
Outline

Introduction

Related work

Context: data and algorithm

Methodology

Results
  - Recommendation performance
  - Gender gaps in recommendations

Adversarial debiasing

Conclusion and perspectives
Related work: Fairness in Recommender Systems

- **Surveys:** Ekstrand et al., 2022; Wang et al., 2022
- **Fairness:** w.r.t. users (our focus), or items (distribution of exposure), or both
- **User fairness:**
  - Are recommendations equally relevant for different groups? Mehrotra et al., 2017; Ekstrand et al., 2018
  - Trade-offs between recommendation performance and other concerns e.g. gender wage gap Rus et al., 2022
  - Causal use of protected variable Kusner et al., 2017
    - Link with audit studies in economics Zhang et al., 2022
- **Algorithmic bias mitigation:** pre / in / post-processing
  - Adversarial in-processing approaches Edwards et al. 2015, Islam et al. 2022, Rus et al. 2022
Outline

Introduction

Related work

Context: data and algorithm

Methodology

Results
  - Recommendation performance
  - Gender gaps in recommendations

Adversarial debiasing

Conclusion and perspectives
Context: Data

- **Scope:** Auvergne-Rhône-Alpes region (France); 2019-mid 2022
- **Dataset size:** 1.2M job seekers, 2.2M job ads, 285k hires
- **Job seeker and job ad characteristics:** both represented in dimension $\sim 500$
  - Include: labor market profile, preferences, background, text vs. wage, labor conditions, required qualifications, text
- **Gender (binary) is available but not used as input**
- **Train-test split:** 85% / 15% on a weekly basis
Context: Algorithm

- **Goal**: rank job ads $y$ for some job seeker $x$
- **Training labels**: hires
- **Two-tiered neural network architecture**: Bied et al., IJCAI 2023
  - Embedding-based first tier (bottom left), designed for scalability, selects 1,000 job ads for each job seeker
  - Second tier (bottom right) re-ranks those using more expressive model / features

![Diagram of two-tiered neural network architecture]

**First tier**

- Job Seeker Variables
  - $\phi_{geo}(x)$
  - $\phi_{sk}(x)$
  - $\phi_{gal}(x)$

- Job Ads Variables
  - $\psi_{geo}(y)$
  - $\psi_{sk}(y)$
  - $\psi_{gal}(y)$

**Second tier**

- Job Seeker Variables and embedding
  - $\phi(x)$

- Job Ads Variables
  - $\psi(y)$

- Pairwise Features
  - $\text{Muse}.0(x, y)$
  - $\text{Muse}.1(x, y)$
  - $\text{Var}(x, y)$

- Muse scores/rank
  - $\text{Muse}.0(x, y) \cdot r.0(x, y)$
Outline

Introduction

Related work

Context: data and algorithm

Methodology

Results
  Recommendation performance
  Gender gaps in recommendations

Adversarial debiasing

Conclusion and perspectives
Methodology (1): Overview

- Is recommendation performance different for men and women?
  - Measure: recall@k, i.e. share of test job seekers s.t. their future hire is in the top \( k \) recommendations

- Are different job ads shown to women and men? In terms of:
  - Wage, distance, executive status, contract type, working hours, male-dominated occupation
  - Fit between to job seeker's search criteria (average fit w.r.t. distance / occupation / wage / contract / working hours)
Methodology (2): Gender Gaps

- Gender $G$ ($=1$ if woman)
- $Y$: characteristic, e.g. wage, of algorithm’s top-1 recommendation
- Naive average recommended quantities:

$$\delta = \mathbb{E}[Y|G = 1] - \mathbb{E}[Y|G = 0]$$

- But: is it reasonable to expect from a “fair” algorithm to disregard job seeker preferences and qualifications?
Methodology (3): Gender Gaps

- $X$: job seeker characteristics used as input
- Let $Z \subset X$ correspond to “job search fundamentals”, which include:
  - Preferences: desired wage, contract type, occupation, accepted mobility
  - Qualifications: education, skills, experience
- Under certain conditions, average difference between genders can be decomposed as (Oaxaca):
  1. An effect explained by job search fundamentals $Z$
  2. And a residual $\tau$ which can not be explained by $Z$
- Main condition: job seekers must be comparable in terms of $Z$
- Statistical model:
  \[ Y = \tau G + \mu_0(Z) + \varepsilon, \quad E(\varepsilon|Z, G) = 0 \]
  where $\mu_0(Z)$ is allowed to be a flexible function.
- $\tau$ is estimated using Double Machine Learning
  
  Chernozhukov et al., 2018
What is the origin of biases (in hirings / recommendations)?

- Job seekers’ biases: when applying to a job, job seekers consider:
  - Chances of being hired → gendered under / over-confidence
  - Utility if hired → gendered valuation of job characteristics, (occupation, wage, distance)

- Recruiters’ biases

Relationship to fairness:

- So far, we have been speaking of biases in a statistical sense
- Reproducing recruiter biases is surely inadmissible wrt fairness
- If biases come from jobseekers, they may or may not be admissible depending on:
  - Origin: gendered job characteristic valuation vs over / under-confidence
  - Chosen normative stance: maximizing job seeker utility vs seeking to reduce labor market inequalities / gender stereotypes
Outline

Introduction

Related work

Context: data and algorithm

Methodology

Results

- Recommendation performance
- Gender gaps in recommendations

Adversarial debiasing

Conclusion and perspectives
Recommendation performance

<table>
<thead>
<tr>
<th>Top $k$</th>
<th>Recall@$k$</th>
<th>Men</th>
<th>Women</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.256</td>
<td>0.243</td>
<td>0.267</td>
<td>0.000</td>
</tr>
<tr>
<td>20</td>
<td>0.351</td>
<td>0.333</td>
<td>0.366</td>
<td>0.000</td>
</tr>
<tr>
<td>100</td>
<td>0.590</td>
<td>0.576</td>
<td>0.603</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Notes: Results on test set hires ($n = 41,787$). Column “p-value” corresponds to a test of equality between columns “Men” and “Women”.

- Recall higher for women than for men
- Difference is statistically significant
- Interpretation attempt: women’s behavior may be easier to predict (mobility? risk aversion?)
## Gender gaps in recommendations

<table>
<thead>
<tr>
<th></th>
<th>Uncond. $\delta$</th>
<th>p-value</th>
<th>Uncond. $\delta$</th>
<th>p-value</th>
<th>Cond. $\tau$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full pop.</td>
<td></td>
<td>Overlap</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wage (log)</td>
<td>-0.023</td>
<td>0.0</td>
<td>-0.016</td>
<td>0.0</td>
<td>-0.004</td>
<td>0.000</td>
</tr>
<tr>
<td>Distance (km)</td>
<td>-0.474</td>
<td>0.0</td>
<td>-0.231</td>
<td>0.0</td>
<td>0.400</td>
<td>0.000</td>
</tr>
<tr>
<td>Executive</td>
<td>-0.004</td>
<td>0.0</td>
<td>-0.009</td>
<td>0.0</td>
<td>-0.002</td>
<td>0.032</td>
</tr>
<tr>
<td>Long term contract</td>
<td>-0.040</td>
<td>0.0</td>
<td>-0.034</td>
<td>0.0</td>
<td>-0.014</td>
<td>0.000</td>
</tr>
<tr>
<td>%Women &lt; 20</td>
<td>-0.411</td>
<td>0.0</td>
<td>-0.219</td>
<td>0.0</td>
<td>-0.033</td>
<td>0.000</td>
</tr>
<tr>
<td>Hours worked per week</td>
<td>-2.934</td>
<td>0.0</td>
<td>-1.957</td>
<td>0.0</td>
<td>-0.381</td>
<td>0.000</td>
</tr>
<tr>
<td>Fit to job search params</td>
<td>-0.028</td>
<td>0.0</td>
<td>-0.019</td>
<td>0.0</td>
<td>-0.011</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Notes: Results on all jobseekers on a test week. Col. 1: average gender gaps $\delta$ ($n = 358,682$). Col. 3: gender gaps $\delta$ for comparable job seekers ($n = 234,145$). Col. 5: gender gap $\tau$ controlling for $Z$ on comparable job seekers.

- Women are, on average, recommended different jobs than men on all selected job characteristics
  - Less paid (2.3 percentage points), less often in executive status, in male-dominated occupations . . .
- The result also holds after controlling for job search fundamentals $Z$, with nevertheless reduced gaps
- In other words, the “unexplained” component of gender gaps is significantly different from 0
### Comparison to application behavior

<table>
<thead>
<tr>
<th>In applications</th>
<th>Differences between women and men</th>
<th>Difference of Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tau_{App}$ (Observed) p-value</td>
<td>$\tau$ (MUSE) p-value</td>
</tr>
<tr>
<td>Wage (log)</td>
<td>-0.012 0.000</td>
<td>-0.011 0.000</td>
</tr>
<tr>
<td>Distance (km)</td>
<td>-4.338 0.000</td>
<td>0.524 0.002</td>
</tr>
<tr>
<td>Executive</td>
<td>-0.002 0.322</td>
<td>-0.002 0.607</td>
</tr>
<tr>
<td>Long term contract</td>
<td>-0.023 0.003</td>
<td>-0.021 0.052</td>
</tr>
<tr>
<td>%Women &lt; 20</td>
<td>-0.142 0.000</td>
<td>-0.067 0.000</td>
</tr>
<tr>
<td>Hours worked/week</td>
<td>-1.177 0.000</td>
<td>-0.675 0.000</td>
</tr>
<tr>
<td>Fit to job search param.</td>
<td>-0.029 0.000</td>
<td>-0.025 0.000</td>
</tr>
</tbody>
</table>

Notes: Results on hired comparable job seekers for whom we observe applications ($n = 12,515$). Col. 1: conditional gender gaps in applications. Col. 3: conditional gender gaps in recommendations. Col. 5: difference of differences, i.e., conditional estimates for the differences between an application's characteristics and the recommendations.

- Gender gaps exist in applications
- The algorithm **does not increase gender gaps**, and reduces some of them (wage, occupation, working hours)
- Same results hold for hiring data
Outline

Introduction

Related work

Context: data and algorithm

Methodology

Results
  Recommendation performance
  Gender gaps in recommendations

Adversarial debiasing

Conclusion and perspectives
Adversarial de-biasing: setup

- **Goal**: de-correlate recommendations from gender
- Algorithm’s first tier (top-1,000 selection) taken as given
- Modify second tier, with adversarial loss:

\[ L_{\text{classif}} - \lambda L_{\text{adv}} \]

where:
- \( L_{\text{classif}} \): BCE loss predicting whether the pair \( i - j \) is a hire
- \( L_{\text{adv}} \): BCE loss of adversary predicting \( i \)'s gender from the latent
## Adversarial de-biasing: results

<table>
<thead>
<tr>
<th>Performance indicators</th>
<th>$\lambda = 0$</th>
<th>p-value</th>
<th>$\lambda = 0.01$</th>
<th>p-value</th>
<th>$\lambda = 1$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R@20</td>
<td>0.351</td>
<td></td>
<td>0.346</td>
<td></td>
<td>0.335</td>
<td></td>
</tr>
<tr>
<td>R@20 (men)</td>
<td>0.333</td>
<td></td>
<td>0.329</td>
<td></td>
<td>0.320</td>
<td></td>
</tr>
<tr>
<td>R@20 (women)</td>
<td>0.366</td>
<td></td>
<td>0.361</td>
<td></td>
<td>0.348</td>
<td></td>
</tr>
<tr>
<td>Adversary’s accuracy</td>
<td></td>
<td></td>
<td>0.784</td>
<td></td>
<td>0.530</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unconditional gaps</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wage (log)</td>
<td>-0.012</td>
<td>0.000</td>
<td>-0.001</td>
<td>0.016</td>
<td>-0.001</td>
<td>0.054</td>
</tr>
<tr>
<td>Distance</td>
<td>0.208</td>
<td>0.043</td>
<td>0.001</td>
<td>0.978</td>
<td>0.046</td>
<td>0.020</td>
</tr>
<tr>
<td>Executive</td>
<td>-0.004</td>
<td>0.028</td>
<td>-0.001</td>
<td>0.132</td>
<td>-0.000</td>
<td>0.273</td>
</tr>
<tr>
<td>Long term contract</td>
<td>-0.051</td>
<td>0.000</td>
<td>-0.011</td>
<td>0.000</td>
<td>-0.011</td>
<td>0.000</td>
</tr>
<tr>
<td>%Women &lt; 20</td>
<td>-0.236</td>
<td>0.000</td>
<td>-0.044</td>
<td>0.000</td>
<td>-0.047</td>
<td>0.000</td>
</tr>
<tr>
<td>Hours worked</td>
<td>-1.939</td>
<td>0.000</td>
<td>-0.340</td>
<td>0.000</td>
<td>-0.313</td>
<td>0.000</td>
</tr>
<tr>
<td>Fit to job search parameters</td>
<td>-0.028</td>
<td>0.000</td>
<td>-0.005</td>
<td>0.000</td>
<td>-0.004</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conditional gaps (DML)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wage (log)</td>
<td>-0.005</td>
<td>0.014</td>
<td>-0.001</td>
<td>0.035</td>
<td>-0.001</td>
<td>0.110</td>
</tr>
<tr>
<td>Distance</td>
<td>0.542</td>
<td>0.000</td>
<td>0.059</td>
<td>0.016</td>
<td>0.100</td>
<td>0.000</td>
</tr>
<tr>
<td>Executive</td>
<td>-0.002</td>
<td>0.319</td>
<td>-0.001</td>
<td>0.177</td>
<td>-0.001</td>
<td>0.052</td>
</tr>
<tr>
<td>Long term contract</td>
<td>-0.027</td>
<td>0.000</td>
<td>-0.005</td>
<td>0.001</td>
<td>-0.006</td>
<td>0.000</td>
</tr>
<tr>
<td>%Women &lt; 20</td>
<td>-0.058</td>
<td>0.000</td>
<td>-0.009</td>
<td>0.000</td>
<td>-0.012</td>
<td>0.000</td>
</tr>
<tr>
<td>Hours worked</td>
<td>-0.695</td>
<td>0.000</td>
<td>-0.103</td>
<td>0.000</td>
<td>-0.132</td>
<td>0.000</td>
</tr>
<tr>
<td>Fit to job search parameters</td>
<td>-0.022</td>
<td>0.000</td>
<td>-0.003</td>
<td>0.000</td>
<td>-0.003</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Notes: Results on hired job seekers, for different weights $\lambda$. Recall and adversary accuracy are computed on the test set ($n = 41,787$). Unconditional and conditional gaps are computed on comparable hired job seekers ($n = 25,783$).

When $\lambda$ increases:

- **R@20 decreases** (0.016 points from $\lambda = 0$ to $\lambda = 1$), esp. for women
- **Adversary’s accuracy drops** (85% when $\lambda = 0.001$, 53% when $\lambda = 1$)
- **Unconditional and conditional gender gaps are considerably reduced**
Outline

Introduction

Related work

Context: data and algorithm

Methodology

Results
  Recommendation performance
  Gender gaps in recommendations

Adversarial debiasing

Conclusion and perspectives
Conclusion

- Recall slightly higher for women than for men

- Gender gaps (conditioned to search fundamentals) exist in recommendations
  - Women’s recommendations are on average paid less, proposed fewer working hours, less often secured by indefinite duration contracts, and less often in male-dominated occupations than men’s

- Same / stronger differences are found in i) actual hiring behavior; ii) application behavior

- Adversarial de-biasing can considerably reduce gender gaps at the expense of recall
Perspectives

- Toward a multi-objective problem: optimize recall (making effective recommendations), comply with js’ preferences (making desirable recommendations) and society’s policy (reducing gaps)

- Required (PES; or EU regulations): specifications about gender gaps (‘not worse than in actual data’; ’better’)

- Caveat: recommendations must be ”sufficiently close” to job seekers’ search (possibly gendered) behavior in order to be considered

- Finding a decent trade-off requires the users’ feedback: focus groups; A/B tests; else?
Datasets used for the analysis

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sample size</th>
<th>Number men</th>
<th>Number women</th>
<th>% men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full week</td>
<td>358,682</td>
<td>176,244</td>
<td>182,438</td>
<td>49.14</td>
</tr>
<tr>
<td>Full week (overlap)</td>
<td>234,145</td>
<td>110,103</td>
<td>124,042</td>
<td>47.02</td>
</tr>
<tr>
<td>Hires</td>
<td>41,787</td>
<td>19,496</td>
<td>22,291</td>
<td>46.66</td>
</tr>
<tr>
<td>Hires (overlap)</td>
<td>25,783</td>
<td>11,434</td>
<td>14,349</td>
<td>44.35</td>
</tr>
<tr>
<td>Hires &amp; Applications (overlap)</td>
<td>12,515</td>
<td>5,517</td>
<td>6,998</td>
<td>44.08</td>
</tr>
</tbody>
</table>
Machine learning algorithm - job seeker features (in $Z$)

<table>
<thead>
<tr>
<th>Preferences</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reservation wage (euros / hour)</td>
<td>numeric</td>
</tr>
<tr>
<td>The job seeker is looking for a full-time job</td>
<td>binary</td>
</tr>
<tr>
<td>Target job sector</td>
<td>categorical (x14)</td>
</tr>
<tr>
<td>Target job</td>
<td>categorical (x110)</td>
</tr>
<tr>
<td>Target type of contract</td>
<td>categorical (x13)</td>
</tr>
<tr>
<td>Maximum commuting time</td>
<td>numeric</td>
</tr>
<tr>
<td>Maximum (and Minimum) number of work hours per week</td>
<td>numeric</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Qualifications</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of years of experience</td>
<td>numeric</td>
</tr>
<tr>
<td>Maximum level of qualification</td>
<td>categorical (x10)</td>
</tr>
<tr>
<td>Department</td>
<td>categorical (x13)</td>
</tr>
<tr>
<td>Vocational training field</td>
<td>categorical (x27)</td>
</tr>
<tr>
<td>Skills (SVD)</td>
<td>numeric (x50)</td>
</tr>
<tr>
<td>Driving licences</td>
<td>categorical (x22)</td>
</tr>
<tr>
<td>Number of languages spoken</td>
<td>numeric</td>
</tr>
<tr>
<td>Means of transportation</td>
<td>categorical (x5)</td>
</tr>
</tbody>
</table>
Machine learning algorithm - job seeker features (not in \(Z\))

<table>
<thead>
<tr>
<th>Socio-demographic variables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of children</td>
<td>numeric</td>
</tr>
<tr>
<td>Jobseeker lives in a QPV area</td>
<td>numeric</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Past employment history</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of unemployment periods in lifetime</td>
<td>numeric</td>
</tr>
<tr>
<td>Reason why the job seeker registered at PES</td>
<td>categorical (x15)</td>
</tr>
<tr>
<td>Type of accompaniment received from PES</td>
<td>categorical (x4)</td>
</tr>
<tr>
<td>Main obstacles assumed to slow return to employment</td>
<td>categorical (x4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resume</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Curriculum text (SVD)</td>
<td>numeric (x100)</td>
</tr>
<tr>
<td>Number of words in the curriculum text</td>
<td>numeric</td>
</tr>
<tr>
<td>Number of visit cards</td>
<td>numeric</td>
</tr>
<tr>
<td>Number of sectors considered by the job seeker</td>
<td>numeric</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geographic information</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Firm density within zip code</td>
<td>numeric</td>
</tr>
<tr>
<td>Unemployment rate within zip code</td>
<td>numeric</td>
</tr>
<tr>
<td>Latitude</td>
<td>numeric</td>
</tr>
<tr>
<td>Longitude</td>
<td>numeric</td>
</tr>
</tbody>
</table>
Comparison to hiring and application behavior: full results

<table>
<thead>
<tr>
<th>In hirings</th>
<th>Differences between women and men</th>
<th>Difference of Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tau_{\text{Hire}}$ (Observed)</td>
<td>$\tau$ (MUSE)</td>
</tr>
<tr>
<td>Wage (log)</td>
<td>-0.010</td>
<td>0.000</td>
</tr>
<tr>
<td>Distance (km)</td>
<td>-1.720</td>
<td>0.022</td>
</tr>
<tr>
<td>Executive</td>
<td>-0.005</td>
<td>0.012</td>
</tr>
<tr>
<td>Long term contract</td>
<td>-0.034</td>
<td>0.000</td>
</tr>
<tr>
<td>%Women &lt; 20</td>
<td>-0.141</td>
<td>0.000</td>
</tr>
<tr>
<td>Hours worked per week</td>
<td>-1.107</td>
<td>0.000</td>
</tr>
<tr>
<td>Fit to job search parameters</td>
<td>-0.019</td>
<td>0.000</td>
</tr>
</tbody>
</table>

| In applications | $\tau_{\text{App}}$ (Observed)  | $\tau$ (MUSE)          | $\tau_{\text{DifA}}$ (MUSE) |
|-----------------|-----------------------------------|----------------------------|
| Wage (log)      | -0.012                            | 0.000                      | 0.002                      | 0.000                      | 0.002 | 0.559 |
| Distance (km)   | -4.338                            | 0.000  | 0.524                      | 0.002                      | 4.905 | 0.000 |
| Executive       | -0.002                            | 0.322  | -0.002                      | 0.607                      | 0.001 | 0.791 |
| Long term contract | -0.023                            | 0.003  | -0.021                      | 0.052                      | 0.002 | 0.900 |
| %Women < 20     | -0.142                            | 0.000  | -0.067                      | 0.000                      | 0.076 | 0.000 |
| Hours worked/week | -1.177                            | 0.000  | -0.675                      | 0.000                      | 0.507 | 0.001 |
| Fit to job search param. | -0.029                            | 0.000  | -0.025                      | 0.000                      | 0.007 | 0.156 |

Notes: Top half: hired job seekers with sufficiently comparable characteristics ($n = 25,783$); bottom half: subset of those for which we also observe applications ($n = 12,515$). First column: conditional gender gaps on hirings (resp. applications). Third columns: conditional gender gaps in recommendations. Fifth column: difference of differences, i.e., the conditional estimates for the differences between a hire’s characteristics (resp application’s) and the recommendations.
Adversarial de-biasing: setup details

- **Goal**: de-correlate recommendations from gender
- Algorithm’s first tier (top-1,000 selection) taken as given
- Modify second tier, with adversarial loss:

\[ L_{classif} - \lambda L_{adv} \]

where:

- \( L_{classif} \): BCE loss predicting whether the pair \( i - j \) is a hire
- \( L_{adv} \): BCE loss of adversary predicting \( i \)'s gender from the latent variables and embedding \( \phi_0(x) \) and \( \psi_0(y) \).
Adversarial de-biasing: full results

<table>
<thead>
<tr>
<th>Performance indicators</th>
<th>( \lambda = 0 )</th>
<th>p-value</th>
<th>( \lambda = 0.001 )</th>
<th>p-value</th>
<th>( \lambda = 0.01 )</th>
<th>p-value</th>
<th>( \lambda = 0.1 )</th>
<th>p-value</th>
<th>( \lambda = 1 )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R@20</td>
<td>0.351</td>
<td></td>
<td>0.346</td>
<td></td>
<td>0.346</td>
<td></td>
<td>0.342</td>
<td></td>
<td>0.335</td>
<td></td>
</tr>
<tr>
<td>R@20 (men)</td>
<td>0.333</td>
<td></td>
<td>0.330</td>
<td></td>
<td>0.329</td>
<td></td>
<td>0.327</td>
<td></td>
<td>0.320</td>
<td></td>
</tr>
<tr>
<td>R@20 (women)</td>
<td>0.366</td>
<td></td>
<td>0.360</td>
<td></td>
<td>0.361</td>
<td></td>
<td>0.356</td>
<td></td>
<td>0.348</td>
<td></td>
</tr>
<tr>
<td>Adversary’s accuracy</td>
<td>0.850</td>
<td></td>
<td>0.784</td>
<td></td>
<td>0.573</td>
<td></td>
<td>0.530</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Unconditional gaps    |                   |         |                   |         |                   |         |                   |         |                   |         |
| Wage (log)            | -0.012            | 0.000   | -0.001            | 0.033   | -0.001            | 0.016   | -0.001            | 0.166   | -0.001            | 0.054   |
| Distance              | 0.208             | 0.043   | -0.003            | 0.882   | 0.001             | 0.978   | 0.040             | 0.050   | 0.046             | 0.020   |
| Executive             | -0.004            | 0.028   | 0.001             | 0.121   | -0.001            | 0.132   | -0.000            | 0.440   | -0.000            | 0.273   |
| Long term contract    | -0.051            | 0.000   | -0.011            | 0.000   | -0.011            | 0.000   | -0.012            | 0.000   | -0.011            | 0.000   |
| %Women < 20           | -0.236            | 0.000   | -0.045            | 0.000   | -0.044            | 0.000   | -0.045            | 0.000   | -0.047            | 0.000   |
| Hours worked          | -1.939            | 0.000   | -0.350            | 0.000   | -0.340            | 0.000   | -0.315            | 0.000   | -0.313            | 0.000   |
| Fit to job search parameters | -0.028 | 0.000   | -0.005            | 0.000   | -0.005            | 0.000   | -0.005            | 0.000   | -0.004            | 0.000   |

| Conditional gaps (DML) |                   |         |                   |         |                   |         |                   |         |                   |         |
| Wage (log)            | -0.005            | 0.014   | -0.001            | 0.109   | -0.001            | 0.035   | -0.000            | 0.281   | -0.001            | 0.110   |
| Distance              | 0.542             | 0.000   | 0.482             | 0.087   | 0.059             | 0.016   | 0.107             | 0.000   | 0.100             | 0.000   |
| Executive             | -0.002            | 0.319   | -0.001            | 0.046   | -0.001            | 0.177   | -0.000            | 0.291   | -0.001            | 0.052   |
| Long term contract    | -0.027            | 0.000   | -0.004            | 0.006   | -0.005            | 0.001   | -0.004            | 0.003   | -0.006            | 0.000   |
| %Women < 20           | -0.058            | 0.000   | -0.009            | 0.000   | -0.009            | 0.000   | -0.011            | 0.000   | -0.012            | 0.000   |
| Hours worked          | -0.695            | 0.000   | -0.105            | 0.000   | -0.103            | 0.000   | -0.111            | 0.000   | -0.132            | 0.000   |
| Fit to job search parameters | -0.022 | 0.000   | -0.003            | 0.000   | -0.003            | 0.000   | -0.003            | 0.000   | -0.003            | 0.000   |

Notes: Results on hired job seekers, for different weights \( \lambda \) given to the adversarial term. Recall and adversary accuracy are computed on the test set (all hired job seekers, \( n = 41,787 \)). Unconditional and conditional gaps are computed on the population of comparable hired job seekers (\( n = 25,783 \)). Unconditional gaps correspond to a difference in means between men and women.