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

Guillaume Bied[1, 2], Christophe Gaillac [3], Morgane Hoffmann [1], Philippe Caillou [2], Bruno Crépon [1], Solal Nathan [2], Michèle Sebag [2]

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

AEQUITAS Workshop at ECAI, September 26th 2023

Outline

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

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

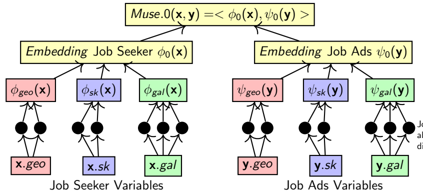
Outline

Context: Data

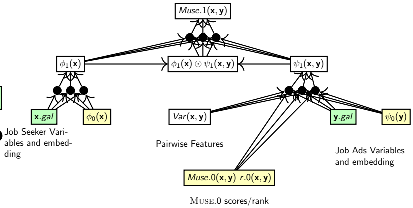
- **Scope:** Auvergne-Rhône-Alpes region (France); 2019-mid 2022
- **Dataset size:** 1.2M job seekers, 2.2M job ads, 285k hires
- **Job seeker and job ad characteristics:** both represented in dimension ~ 500
  - Include: labor market profile, preferences, background, text vs. wage, labor conditions, 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 $\mathbf{y}$ for some job seeker $\mathbf{x}$
- **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



First tier



Second tier

Outline

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 τ 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.
- τ is estimated using *Double Machine Learning*

Chernozhukov et al., 2018

Discussion

What is the origin of biases (in hirings / recommendations)?

- Job seekers' biases: when applying to a job, job seekers consider:
  - Chances of being hired $\rightarrow$ gendered under / over-confidence
  - Utility if hired $\rightarrow$ 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

Recommendation performance

Top k	Recall@k	Men	Women	p-value
10	0.256	0.243	0.267	0.000
20	0.351	0.333	0.366	0.000
100	0.590	0.576	0.603	0.000

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

	Uncond. δ *Full pop.*	p-value	Uncond. δ *Overlap*	p-value	Cond. τ	p-value
Wage (log)	-0.023	0.0	-0.016	0.0	-0.004	0.000
Distance (km)	-0.474	0.0	-0.231	0.0	0.400	0.000
Executive	-0.004	0.0	-0.009	0.0	-0.002	0.032
Long term contract	-0.040	0.0	-0.034	0.0	-0.014	0.000
%Women < 20	-0.411	0.0	-0.219	0.0	-0.033	0.000
Hours worked per week	-2.934	0.0	-1.957	0.0	-0.381	0.000
Fit to job search parameters	-0.028	0.0	-0.019	0.0	-0.011	0.000

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

In applications	Differences between women and men τ_{App} (Observed)	p-value	τ (MUSE)	p-value	Difference of Differences τ_{DifA} (MUSE)	p-value
Wage (log)	-0.012	0.000	-0.011	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: 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 recommmendations. 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

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:

$$L_{classif} - \lambda L_{adv}$$

  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

Adversarial de-biasing: results

	$\lambda = 0$	p-value	$\lambda = 0.01$	p-value	$\lambda = 1$	p-value
Performance indicators						
R@20	0.351		0.346		0.335	
R@20 (men)	0.333		0.329		0.320	
R@20 (women)	0.366		0.361		0.348	
Adversary's accuracy			0.784		0.530	
Unconditional gaps						
Wage (log)	-0.012	0.000	-0.001	0.016	-0.001	0.054
Distance	0.208	0.043	0.001	0.978	0.046	0.020
Executive	-0.004	0.028	-0.001	0.132	-0.000	0.273
Long term contract	-0.051	0.000	-0.011	0.000	-0.011	0.000
%Women $<$ 20	-0.236	0.000	-0.044	0.000	-0.047	0.000
Hours worked	-1.939	0.000	-0.340	0.000	-0.313	0.000
Fit to job search parameters	-0.028	0.000	-0.005	0.000	-0.004	0.000
Conditional gaps (DML)						
Wage (log)	-0.005	0.014	-0.001	0.035	-0.001	0.110
Distance	0.542	0.000	0.059	0.016	0.100	0.000
Executive	-0.002	0.319	-0.001	0.177	-0.001	0.052
Long term contract	-0.027	0.000	-0.005	0.001	-0.006	0.000
%Women $<$ 20	-0.058	0.000	-0.009	0.000	-0.012	0.000
Hours worked	-0.695	0.000	-0.103	0.000	-0.132	0.000
Fit to job search parameters	-0.022	0.000	-0.003	0.000	-0.003	0.000

Notes: Results on hired job seekers, for different weights λ. 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 λ 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

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

	Sample size	Number men	Number women	% men
Full week	358,682	176,244	182,438	49.14
Full week (overlap)	234,145	110,103	124,042	47.02
Hires	41,787	19,496	22,291	46.66
Hires (overlap)	25,783	11,434	14,349	44.35
Hires & Applications (overlap)	12,515	5,517	6,998	44.08

Machine learning algorithm - job seeker features (in Z)

Preferences	
Reservation wage (euros / hour)	numeric
The job seeker is looking for a full-time job	binary
Target job sector	categorical (x14)
Target job	categorical (x110)
Target type of contract	categorical (x13)
Maximum commuting time	numeric
Maximum (and Minimum) number of work hours per week	numeric
Qualifications	
Number of years of experience	numeric
Maximum level of qualification	categorical (x10)
Department	categorical (x13)
Vocational training field	categorical (x27)
Skills (SVD)	numeric (x50)
Driving licences	categorical (x22)
Number of languages spoken	numeric
Means of transportation	categorical (x5)

Machine learning algorithm - job seeker features (not in Z)

Socio-demographic variables	
Number of children	numeric
Jobseeker lives in a QPV area	numeric
Past employment history	
Number of unemployment periods in lifetime	numeric
Reason why the job seeker registered at PES	categorical (x15)
Type of accompaniment received from PES	categorical (x4)
Main obstacles assumed to slow return to employment	categorical (x4)
Resume	
Curriculum text (SVD)	numeric (x100)
Number of words in the curriculum text	numeric
Number of visit cards	numeric
Number of sectors considered by the job seeker	numeric
Geographic information	
Firm density within zip code	numeric
Unemployment rate within zip code	numeric
Latitude	numeric
Longitude	numeric

Comparison to hiring and application behavior: full results

	Differences between women and men				Difference of Differences	
In hirings	τ_{Hire}(*Observed*)	p-value	τ (MUSE)	p-value	τ_{DifH} (MUSE)	p-value
Wage (log)	-0.010	0.000	-0.005	0.014	0.004	0.099
Distance (km)	-1.720	0.022	0.542	0.000	2.196	0.003
Executive	-0.005	0.012	-0.002	0.319	0.003	0.365
Long term contract	-0.034	0.000	-0.027	0.000	0.008	0.442
%Women < 20	-0.141	0.000	-0.058	0.000	0.084	0.000
Hours worked per week	-1.107	0.000	-0.695	0.000	0.441	0.001
Fit to job search parameters	-0.019	0.000	-0.022	0.000	-0.002	0.557
In applications	τ_{App} (Observed)	p-value	τ (MUSE)	p-value	τ_{DifA} (MUSE)	p-value
Wage (log)	-0.012	0.000	-0.011	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 recommmendations. Fifth column: difference of differences, *i.e.*, the conditional estimates for the differences between a hire's characteristics (resp application's) and the recommendations.

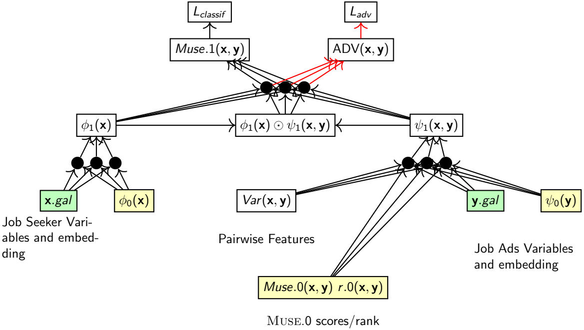
Adversarial de-biasing: setup details

- **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



Adversarial de-biasing: full results

	$\lambda = 0$	p-value	$\lambda = 0.001$	p-value	$\lambda = 0.01$	p-value	$\lambda = 0.1$	p-value	$\lambda = 1$	p-value
Performance indicators										
R@20	0.351		0.346		0.346		0.342		0.335	
R@20 (men)	0.333		0.330		0.329		0.327		0.320	
R@20 (women)	0.366		0.360		0.361		0.356		0.348	
Adversary's accuracy			0.850		0.784		0.573		0.530	
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 λ 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.