

# Fairness in job recommendations: estimating, explaining, and reducing gender gaps<sup>\*</sup>

Guillaume Bied<sup>1,2,\*</sup>, Christophe Gaillac<sup>3</sup>, Morgane Hoffmann<sup>4</sup>, Philippe Caillou<sup>1</sup>, Bruno Crépon<sup>2</sup>, Solal Nathan<sup>1</sup> and Michèle Sebag<sup>1</sup>

<sup>1</sup>Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), Orsay, France

<sup>2</sup>Centre de Recherche en Economie et Statistique (CREST), Palaiseau, France

<sup>3</sup>Nuffield College and Oxford University, Oxford, United Kingdom

<sup>4</sup>Pôle emploi, Paris, France

## Abstract

Algorithmic recommendations of job ads have the potential to reduce frictional unemployment, but raise concerns about fairness due to biases in past data. Our research investigates the issue of algorithmic fairness with a specific focus on gender in a hybrid job recommendation system developed in partnership with the French Public Employment Service (PES), which is trained on past hires. First, by viewing job ads as a set of characteristics (such as wage and contract type), we document how the algorithm treats job seekers differently based on gender, both unconditionally and conditionally on their search parameters and qualifications. Second, we discuss the notion(s) of algorithmic fairness applicable in this context and the trade-offs involved. We show that the considered system reflects some existing differences in hiring or applications but does not exacerbate them. Finally, we consider adversarial de-biasing technique as a practical tool to demonstrate the trade-offs between recall and reduced differentiated treatment.

## Keywords

Fairness, Job recommender systems, Adversarial de-biasing, Gender gaps, Human resources

## 1. Introduction

At the core of e-business, recommender systems leverage past data to help users locate relevant items among large amounts of possible ones that would be costly to explore otherwise. Since an important part of unemployment can be explained by informational frictions [1], including the costs of acquiring information and cognitive limitations, recommender systems could improve matching on the labor market. As labor market outcomes shape livelihoods, social positions and individual identities, helping job seekers find the right jobs matters.

Yet job recommender systems are also a textbook case of fairness issues in machine learning [2]. Algorithms trained on real-world data, which involve human biases and discriminatory practices, may reproduce, or even increase, past undesirable behavior such as gender stereotypes,

---

*Aequitas 2023: Workshop on Fairness and Bias in AI | co-located with ECAI 2023, Kraków, Poland*


<sup>\*</sup>The present work represents the views of its authors. It does not represent the views or opinions of the French PES. The algorithm audited in the present work is not currently used by the PES to issue recommendations to job seekers.

\*Corresponding author.

✉ [bied@lri.fr](mailto:bied@lri.fr) (G. Bied)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

and widen labor market inequalities. Ensuring this does not happen is a major concern for the scientific community, Public Employment Services as well as for all citizens.

This paper investigates the issue of gender fairness within the context of the audit of a recommender system called *MULTI-head Sparse E-recruitment* (MUSE hereafter) [3], developed in partnership with the French Public Employment service (PES). MUSE leverages extensive data about job seekers' and job ads' characteristics and learns from past hiring patterns. Our contributions are threefold. Firstly, we discuss the appropriate notion of algorithmic fairness that should be adopted in the PES setting. Gender disparities in hirings, viewed in terms of job characteristics, such as occupation, distance, wage, full or part-time status can arise from differentiated application choices arising from job seekers' preferences. The algorithm's replication of this behavior appears justified in maximizing users' welfare (see the related individual, envy-freeness, and preference-based notions of fairness respectively in [4, 5, 6]). However, these gaps can also arise from differential valuations of inherent job seeker's characteristics by recruiters based on gender, which can be seen as discriminatory or unfair. Secondly, we propose to disentangle the impact of job search fundamentals (search parameters and qualifications) from other job seekers' characteristics in explaining observed gaps by using double machine learning [7]. We analyse job ad recommendations and document gender disparities both unconditionally and conditionally on job search fundamentals, showing that these standalone do not fully account for the observed gender gaps. Nevertheless, the system does not exacerbate existing differences in hiring or applications. This discussion brings forth a tension between a PES's missions and values: providing optimal person-dependent recommendations regarding access to employment while ensuring fair treatment between women and men. Finally, we illustrate this trade-off by developing an adversarial de-biasing approach [8] aiming at making recommendations gender-blind. Although this approach reduces differential treatment, it also leads to an overall performance loss and reduction in access to employment, which is more pronounced for women.

The rest of the paper is structured as follows. Section 2 describes the data and the MUSE algorithm. Section 3 proposes to leverage the Double Machine Learning method (DML hereafter) [7, 9, 10] to make inference on the effect of gender on the recommendations, while controlling for the channel of the job search fundamentals. Section 4 audits the algorithm in terms of recommendation performance, provides evidence of differentiated treatment, and compares these differences to those found in hiring and application behavior. Section 5 introduces adversarial techniques to reduce recommendation reliance on gender, and documents their impact on performance metrics and differentiated treatment. Section 6 concludes and provides perspectives for further work. Appendix D contains a simple model explaining the different potential sources of differential treatment and relating them to gender inequalities in observed applications and hires.

**Related work.** Fairness in the context of recommender systems draws an increasing amount of work, surveyed by [11, 12, 13]. Depending on the application domain, fairness issues may arise w.r.t. items (sharing users' attention in an equitable way), w.r.t. users (presenting a fair selection of items to the users), or both [14, 15, 16]. In the present work, we focus on user fairness.

Some approaches to user fairness question whether recommendations are equally relevant for different groups of users in terms of standard metrics such as recall or NDCG. [17] audits search engines for differential satisfaction between demographics. [18] extends this investigation to several public recommendation datasets, discussing whether different groups of users (in terms of age or gender) retrieve the same utility from recommendations based on standard metrics. Such differences may be due to class imbalance, which may lead a recommender system to better capture the interaction patterns of a majority group in a collaborative filtering setting [19]. [20] measure fairness both in terms of differentiated values of predicted ratings conditionally on characteristics, as well as wrt prediction errors between genders.

Other works emphasize the trade-off between recommendation performance and other fairness measures. Among them, [21] approach the problem of collaborative filtering under the lenses of a notion of neutrality akin to demographic parity: recommendations should not vary according to a user-specified viewpoint such as gender. However, with labor market applications in mind, [20] argue that such metrics possibly ignore some legitimate links between gender and preferences. In a labor market context, [22] is concerned with occupation recommendation while reducing the gender wage gap. [23] conduct a correspondence study of several Chinese job boards, demonstrating that some profiles are recommended different job ads depending on whether they are labelled women or men, thus showing a significant causal impact of gender.

Finally, several approaches exist to prevent fairness issues: pre-processing, in-processing and post-processing. Adversarial in-processing methods, initially proposed in the classification setting [8, 24, 25], attempt to decorrelate neural representations with gender. The approach has been proposed for neural recommenders in a labor market setting [22, 20, 26] with different motivations and notions of fairness in mind.

## 2. Experimental setting

**Overview of the data** The proprietary dataset provided by the French PES contains characteristics of job ads and registered job seekers, as well as their interactions, from 2019 to mid-2022 in the Auvergne-Rhône-Alpes region.

The  $i$ -th job seeker’s characteristics, represented as a vector  $x_i \in \mathbb{R}^{483}$  after pre-processing, include job search criteria, labor market profile information, and administrative data (see Appendix B for more details). Within  $x_i$ , job seekers’ search fundamentals (search criteria and qualifications, denoted  $z_i$ ) include desired wage, occupation, geographic location and accepted mobility, search for a full-time or part-time job, qualification level of the desired position, and accepted working hours.

Overall, the labor market profile information in  $x_i$  includes experience, hard and soft skills provided in the PES’s ontology, possession of a driver’s license, educational achievements, textual data (CV, description of past work experience), and administrative data (number of past unemployment spells, reasons for registration, and the type of follow-up provided by the PES). Skills and textual descriptions are each reduced by singular value decomposition [27]. It is emphasized that job seekers’ gender is available as a binary variable, although it is *not* provided to the recommender system.

Similarly, the  $j$ -th job ad is represented by vector  $y_j \in \mathbb{R}^{469}$  after pre-processing. Available

features include lower and upper bounds for the offered wage, workplace postcode, desired skills, requirements in terms of education, contract type, working hours, and textual descriptions of the firm and position. Textual information and skills are also reduced by singular value decomposition. We also observe whether a job seeker  $i$  applied to a job ad  $j$ , and whether he or she was hired on that position. The train and test set cover 1.2 million job seekers and 2.2 million job ads. The 285,992 observed hires are split between train and test on a weekly basis: 85% of weeks are assigned to the train set (representing 241,715 hires), and the rest to the test set (44,277 hires).

**Datasets used for the analysis** The algorithm’s recommendations will be studied using several distinct datasets.

To study gender gaps conditional on job seeker search fundamentals (more in section 3), we restrict the analysis to men and women that cannot be perfectly distinguished on the basis of their characteristics, following the *overlap / common support* assumption [28]. More precisely, if individuals’ gender could be accurately predicted on the basis of characteristics, one could hardly disentangle the impact of such characteristics and that of gender on the recommendations.

The population with common support is selected as follows. The prediction of gender is achieved using random forest [29] considering selected features including education, desired wage, experience, geographic location, desired contract type, occupation, level of qualification, search for part-time job, accepted mobility. The learned classifier, referred to as *propensity score*, with accuracy circa 88% is used to select the job seekers in the common support, retaining individuals with propensity score in  $[0.05, .95]$ .

To study recommendations issued to all job seekers at a given point in time, we consider all job seekers registered during a randomly chosen week of the test set (the fourteenth ISO week of 2022). In order to measure recommendation performance, and to contrast differentiated treatment by the algorithm with differences observed in hiring behavior, we also consider recommendations to all job seekers which are hired during the test weeks. To study application behavior, we consider the average characteristics (all weeks pooled together) of the applications of job seekers for which hires are observed in the test set (we observe 169,325 such applications after restriction to the common support).

The sizes, compositions in terms of gender, and size after restriction to job seekers in the common support, of the datasets of interest are reported in Table 4 in Appendix.

## 2.1. Algorithm

The algorithm MUSE is briefly described for the sake of self-containedness, referring the reader to [3] for a more comprehensive presentation.

**Architecture** MUSE is a two-tier hybrid job recommender system, designed to address the sparsity and cold start issues inherent to the job recommendation setting, and to meet computational requirements. It is trained on hiring data. Hires, rather than other type of interactions, are chosen as training labels since they indicate strong mutual interest of the job seeker and recruiter.

The first tier of the algorithm aims at retrieving a subset of 1,000 job ads (to be re-ranked by the second tier) efficiently. It is a two-tower model, trained with a triplet margin loss, which constructs embeddings for job seekers and job ads based on their contextual information  $x_i$  and  $y_j$ . It correctly keeps 82.25% of matches in the test set among its top-1,000 selection. In the following, we take this first stage operation as given, discarding all job ads but those ranked among the top 1,000 for each job seeker.

The second tier of the algorithm takes as input: the job seeker’s description  $x_i$ ; the job ad’s description w.r.t.  $i$ -th job seeker, noted  $y_{ij}$ , formed of the job ad description  $y_j$  concatenated with the score and rank of associated to job ad  $j$  for  $i$  by the first tier of the algorithm, and the distance in kilometers between  $i$  and  $j$ . Two embeddings respectively denoted  $\phi$  and  $\psi$  are learned on the top of  $x_i$  and  $y_{i,j}$ ; with  $l_{ij}$  formed as the concatenation of these embeddings and their element-wise product ( $l_{ij} = [\phi(x_i), \psi(y_{ij}), \phi(x_i) \odot \psi(y_{ij})]$ ). The recommendation score  $\hat{m}_{ij}$  is learned as a standard neural net on the top of  $l_{ij}$ :

$$\hat{m}_{ij} = f_{\beta}(l_{ij})$$

where  $f_{\beta}$  is a one-hidden layer feedforward neural network parameterized by  $\beta$ . Model parameters are learned end-to-end with a cross-entropy loss:

$$\min_{\beta, \phi, \psi} L := \sum_{i,j} m_{ij} \log(\hat{m}_{ij}) + (1 - m_{ij}) \log(1 - \hat{m}_{ij}),$$

where  $m_{ij}$  is 1 iff  $j$  hired  $i$  and 0 otherwise. In practice, negative examples (pairs which are not matches) are sampled uniformly at random within the first tier’s top-1,000 selection. To issue recommendations, job ads are ranked by decreasing  $\hat{m}_{ij}$ .

### 3. Measuring the effect of gender on recommendations controlling for the preferences

#### 3.1. Measures of interest

We seek to measure how the algorithm’s recommendation performance varies between men and women, but also how the characteristics of the job ads depend on gender, unconditionally and conditionally on job search fundamentals.

**Recommendation performance** will be measured by the recall@ $k$ , defined as the share of hires correctly ranked among the algorithm’s top  $k$  recommendations in the test set.

**Characteristics of recommended jobs** We study gendered differences in terms of the following characteristics of the top recommended job ad: 1) The logarithm of the ad’s wage; 2) The distance in kilometers of the job’s workplace to the job seeker’s zip code; 3) Whether the job ad corresponds to an executive position in the company; 4) Whether the contract is defined for an indefinite duration or not; 5) The number of hours worked per week; 6) Whether the share of women among job seekers searching for a job in the occupation is less than 20%.

We also consider an aggregate indicator of the fit between the job seeker’s search criteria and the recommended job,<sup>1</sup> defined as an average of five binary indicators describing the fit w.r.t. to the job seeker’s i) accepted geographic mobility; ii) desired type of occupation; iii) desired wage; iv) desired type of contract; v) desired working hours.

### 3.2. Methodology

**Parameters of interest.** We seek to document whether different jobs are recommended to women and men on average and conditionally on their job search fundamentals (search parameters and qualifications). Previous studies in economics [30] have documented gendered preferences for commuting time, contract type and wage. These preferences are reflected in jobseekers’ job search parameters and may partially explain observed gender dissimilarities in recommendations. However, the different job search parameters might not be the only ones contributing to the differences in recommendations: the study will try to identify whether other gendered features have an impact on recommendations. Our method disentangles disparities due to different job search fundamentals from those due to other characteristics and their valuation by the algorithm.

If disparities due to preferences are potentially justifiable from the users’ perspective, the other ones could be considered as a sign of unfair algorithmic treatment.

In the following, the covariate  $X$  stands for the whole set of variables describing the job seekers, and used by the recommendation system; it includes information on past employment history, demographics (e.g number of children), and self description in the text of the resume. The control  $Z$  stands for the variables describing the job search fundamentals (job search parameters and qualifications, detailed in Appendix B;  $Z \subset X$ ). The outcome  $Y$  of the recommendation system includes the set of variables describing the recommended job ad (job type, wage, whether the job is part-time or full-time) and cross-features (distance between the locations of the job seeker and the job ad, fit w.r.t. the job seeker’s aggregated search criteria).

The question of gender-related bias arises when men and women with same search fundamentals  $Z$  are recommended substantially different job ads (different outcomes  $Y$ ): even though the system has no direct access to the gender  $G$ , it might value the characteristics in  $X - G$  in a gender-biased way.

To assess this potential differential treatment, we focus on two quantities separately. First, we consider the naive average characteristics  $Y$  of the recommended offers:

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

$G = 1$  and  $G = 0$  denoting respectively women and men hereafter. This parameter can simply be estimated by taking difference in means. Following our discussion, it is questionable whether it is the role of a (fair) recommender system to directly disregard job search fundamentals  $Z$ . Our parameter of interest is thus the gender related gap  $\tau$  in one recommended job ads characteristic  $Y$ , while controlling for the effects of  $Z$ . Taking inspiration of [31, 9] regarding

---

<sup>1</sup>This indicator is inspired from the proprietary PES indicator used for querying job ads.

the estimation of the gender wage, we thus consider the following model:<sup>2</sup>

$$Y = \mu_0(Z) + \tau G + \varepsilon, \quad \mathbb{E}(\varepsilon|G, Z) = 0, \quad (1)$$

where  $\mu_0(z) := \mathbb{E}(Y|G = 0, Z = z)$  are the expected characteristics of jobs for men with preferences  $Z$ , and  $\varepsilon$  is a noise variable. To be able to identify  $\tau$ , we also make the standard assumption of common support, stating that there exists both men and women sharing all types of search parameters  $Z$ , *i.e.*, for all  $z$ , there exists  $\epsilon > 0$ , s.t  $p(z) := \mathbb{P}(G = 1|Z = z) \in [\epsilon, 1 - \epsilon]$ .

Let us give more intuition about the interpretation of the effect  $\tau$  in our context of the impact of gender on recommendations. Consider a linear specification of the effect of the different job search parameters on the recommendations in (1), *i.e.*,  $\mu_0(Z) = Z'\gamma_0$ . Here, denoting by  $\gamma_1$  and  $\gamma_0$  the coefficients of the regression of  $Y$  on  $Z$  for women and men respectively, we obtain the *Oaxaca decomposition*, used in the literature on gender wage gap [31, 9], of the average effect:

$$\delta = \underbrace{\gamma_0'(\mathbb{E}(Z|G = 1) - \mathbb{E}(Z|G = 0))}_{\text{Explained effect by } Z} + \underbrace{(\gamma_1 - \gamma_0)'\mathbb{E}(Z|G = 1)}_{=\tau, \text{ unexplained effect}},$$

where  $\tau$  is the residual of the average gender difference  $\delta$  that cannot be explained by  $Z$ .

Estimation of the *gender gap*  $\tau$  is performed using the double machine learning method (DML) [see, *e.g.*, 7, 10]. This methods provides an estimator of  $\tau$  which is asymptotically normal, robust to the preliminary estimation of other *nuisance* parameters,  $m(Z) := \mathbb{E}(Y|Z)$  and the propensity score  $p(Z) := \mathbb{P}(G = 1|Z)$  using different machine learning estimators. Details are given in Appendix C.

## 4. Results

In all the tables presented hereafter in this section, the column “p-value” presents the p-value indicating the significance of the measure reported in the adjacent left column. Results presented in this section use random forest estimators for functions  $m$  and  $p$ . However, our results are not sensitive to the choice of the estimator as shown in Appendix E.

### 4.1. Recommendation performance is higher for women

We first report the  $\text{recall}@k$  for all hires in the test set, as well as for male and female job seekers separately. For instance, the algorithm correctly ranks within its top 20 (resp. top 50) recommendations the job ad on which a job seeker was hired in 35% (resp. 49%) of cases. This success rate is 33.3% for men (resp. 47.5), and 36.6% for women (resp. 50%), with a statistically significant difference (more on Table 5, Appendix). More generally, we find the  $\text{recall}@k$  to be higher for women than for men at all values of  $k$  considered. While the magnitude of the difference is limited, it is statistically significant. The observed higher performance of the algorithm for women could be explained by the importance given by the model to the distance criterion. Women assign greater value to proximity when searching for a job, see Table 2 on hires and applications, which could make their job choices easier to predict.

<sup>2</sup>The presented methodology follows the CATE identification procedure [see, 28], being granted that the gender cannot be considered a treatment.

## 4.2. Characteristics of job ads recommended to men and women are different

Table 1 provides conditional and unconditional estimates for gender differences in recommended offer characteristics for all registered job seekers and the selected sub-population (section 2).

The first and third columns show that, whatever the restrictions on the population, women are on average recommended different jobs than men. Their recommended job ads are paid 2.3% less than men; half a kilometer closer to home, shorter in terms of weekly working hours (by 2.9 hours); less often of indefinite duration (4 percentage points less often), and executive status (0.4 percentage points). Recommended jobs are also less often in male-dominated occupations (41% less often). Women’s recommended jobs also have a lesser degree of fit with their own search criteria (a loss of 0.028 points in the aggregate fit measure between 0 and 1). All of these differences are statistically significant.

However, the results using the DML estimation (Table 1, column Cond.  $\tau$ ) show that restricting the analysis to the population of job seekers with common support and conditioning on job seeker’s search fundamentals  $Z$  leads to a reduced gender gap in all discussed job ads characteristics. Nevertheless, after conditioning on  $Z$ , women’s recommended jobs still fit less with their search parameters (by 0.011 points), and remain significantly different in all discussed dimensions. For instance, 17% of the wage gender gap is left unexplained by job search characteristics and qualifications of job seekers.

	Uncond. $\delta$ <i>Full pop.</i>	p-value	Uncond. $\delta$ <i>Overlap</i>	p-value	Cond. $\tau$	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: The first column reports the gender gap  $\delta$  in terms of job characteristics on average. The third column reports the gender gap on the population of job seekers with a propensity score between 0.05 and 0.95. The fifth column reports, on the population of job seekers with sufficiently comparable characteristics, the estimates for the gender gap  $\tau$  controlling for search parameters using DML. Results are given using random forests as estimators for the functions  $m$  and  $p$  and are robust to this choice as shown in Appendix E.

**Table 1**

**Unconditional and conditional gender differences in characteristics of offers recommended**

## 4.3. Inequalities in recommendations against the ones observed in hiring and applications

**Inequalities in recommendations are comparable or smaller than the observed ones in hirings.** We turn to the comparison of the characteristics of the recommended job ads to those observed in real-world hires ( $\tau_{\text{Hire}}$ ). We focus on the job seekers in the test set for which we observe hires.



The first column of the upper section of Table 2 shows that, for the population with common support and conditionally on job seeker’s search criteria, there exist differences in hiring behavior  $\tau_{\text{Hire}}$  between women and men. Women are hired on job ads that have a lower aggregate fit (by 0.019) with their search criteria than men. They are hired less often in male-dominated occupations (14.1pp); are less often hired on indefinite duration duration contracts (3.4pp), and work less hours (1.11 hours). All of these differences are statistically significant. Moreover, they are hired on jobs that are paid less,<sup>3</sup> and are less often hired in executive positions.

On the other hand, the third column of Table 2, which reports estimates for  $\tau$  in recommendations for the subsample of hired job seekers, illustrates that the patterns are similar to those established on the whole population in the fifth column of Table 1. However, the gap between the characteristics of hires and the characteristics of recommended job ads after conditioning on  $Z$  ( $\tau_{\text{DifH}}$ ) presented in the fifth column of Table 2 show that they are somehow comparable. Indeed, the algorithm has little impact on the fit between job seeker’s search criteria and the job ads, and does not increase the gap in wages, executive status or long term contracts. Surprisingly, the algorithm seems to recommend job ads in occupations where men are over-represented less often, and recommends positions with more working hours, thus slightly reducing gender gaps.

Eventually, if the algorithm recommends different types of offers to men and women, there is no evidence that it increases the inequalities already observed on the labor market when we condition for job seekers’ job search fundamentals.

**Observed differences largely replicate those observed in application behavior.** Differential treatment in hires may originate from job seekers’ application behavior and from recruiters’ discriminatory behavior (see, *e.g.*, the formal model in appendix D). In the present section, we wish to compare the magnitude of the gender gap in the algorithm’s recommendations  $\tau$  to the magnitude of the gender gap found in job seekers’ applications  $\tau_{\text{App}}$ . As applications can also be seen as a noisy proxy for job seekers’ utility, especially if application costs are low (see appendix D), if the differences  $\tau_{\text{DifA}}$  were large, this would indicate that the algorithm’s learned recommendations reflect job seekers’ preferences but also recruiter biases.

Due to different data sources, we study the sub-population of job seekers with hires in the test weeks for which we observe applications (all weeks pooled together).

The first column of the second panel of Table 2 reports estimates for gender gaps  $\tau_{\text{App}}$  in applications conditionally on  $Z$ . Indeed, the conditional estimates for the gender gaps are significant in application behavior, in terms of fit to search criteria (a significant difference of 0.029 points in the aggregate index), wages, long term contracts, full time jobs, weekly working hours, and occupations where men are over-represented.

Crucially, based on results on the fifth column of the second panel of Table 2, the conditional estimate for the difference between applications’ characteristics and the algorithm’s recommendations  $\tau_{\text{DifA}}$  are not statistically significantly different from zero with respect to fit to search criteria and to all objective job characteristics aside from occupations where men

---

<sup>3</sup>An estimate of 1% for the gender wage gap on the job offers, conditional on search criteria, might be surprising considering the larger magnitudes generally discussed in the economics literature. It should be noted that we have a large set of stated preferences and that the analysis focuses on registered job seekers (rather than on the working population as a whole), with jobs closer to the national minimum wage than those in the national population.

In hirings	Differences between women and men				Difference of Differences	
	$\tau_{\text{Hire}}(\text{Observed})$	p-value	$\tau$ (MUSE)	p-value	$\tau_{\text{DiffH}}$ (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	$\tau_{\text{App}}$ (Observed)	p-value	$\tau$ (MUSE)	p-value	$\tau_{\text{DiffA}}$ (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 are presented on the subsample of hired job seekers. Due to different data sources, we study the sub-population of job seekers with hires in the testing weeks for which we observe applications (all weeks taken together). The first column presents the conditional estimates for the gender gaps on observed hirings (resp. observed applications) between women and men for the population with common support. The third one presents the same difference on the characteristics of the algorithm’s recommendations. For hirings, differences with DML effects presented in the fifth column of Table 1 are due to the restriction on the subsample of hired job seekers. The fifth column reports the difference of two latter differences, *i.e.*, the conditional estimates for the differences between a hire’s characteristics (resp application’s) and the algorithm’s recommendation.

**Table 2**  
**Conditional gender gaps in *hirings* and *applications* and in the algorithm’s recommendations on the subsample of hired job seekers**

are over-represented and number of hours worked. In the two latter cases, the differences in conditional gender gaps is reduced in the algorithm’s recommendations.

Altogether, gender gaps exist in the algorithm’s recommendations even after conditioning on job seekers’ search fundamentals, but those gaps are not larger than those found in hires or in job seekers’ application behavior. These results suggest that the recall, the relevance w.r.t. job seekers’ search fundamentals, and the reduction of the gender-related gaps in recommendation might be antagonistic.

This conjecture will be investigated empirically using adversarial techniques in the next section.

## 5. Limiting differential treatment with adversarial methods

The goal of this section is to investigate the consequences of de-correlating the latent representations from the gender  $g_i$ , using an adversarial method [24, 8, 22, 20], in terms of gender gaps and recall.

## 5.1. Methodology: gender-blind recommendation through adversarial learning

In the following, we take the pre-selection of 1,000 job ads by the first tier of the algorithm as given (considering job ads ranked beyond 1,000 to be irrelevant), and incorporate the adversarial setup to the second tier of the recommender system. Recall that in the usual setting, the algorithm minimizes:

$$\min_{\theta, \beta} L_{classif} := \sum_{i,j} m_{ij} \log(\hat{m}_{ij}) + (1 - m_{ij}) \log(1 - \hat{m}_{ij}),$$

where  $\theta$  corresponds to the weights parameterizing the latent representation  $l_{ij}$  of job seeker  $i$  and job ad  $j$  (viewed with respect to its relation to  $i$ ). The adversary is instantiated as a three-hidden-layer feedforward neural network predicting gender from the latent. Denote its prediction for gender by  $\hat{g}_{ij} = g_{\zeta}(l_{ij})$ , the adversary then tries to solve:

$$\min_{\zeta} L_{adv} = \sum_{i,j} g_i \log(\hat{g}_{ij}) + (1 - g_i) \log(1 - \hat{g}_{ij}),$$

whereas the recommender system incurs a penalty if the adversary's predictions perform well, leading to the program:  $\min_{\theta, \beta} L_{classif} - \lambda L_{adv}$ , where  $\lambda > 0$  is a hyper parameter prioritizing amongst the two objectives. In practice, we alternate between stochastic gradient updates of the two sets of parameters  $\{\zeta\}$  and  $\{\theta, \beta\}$ .

## 5.2. Results

Table 3 presents the performance, unconditional and conditional gender gaps associated to recommendations obtained using the adversarial strategy, letting  $\lambda$  range over  $\{0.001, 0.01, 0.1, 1\}$ .

Adopting the adversarial penalization strategy leads to a slight loss in recall@20: a difference of 0.016 points between  $\lambda = 0$  and  $\lambda = 1$ . While recall remains higher for women than for men, women bear most of the loss (0.018 points, against 0.013 for men) due to adversarial de-biasing (see a theory about this risk in [32]). As  $\lambda$  increases, the gender predictions made by the adversary become less accurate (the accuracy drops from 85% when  $\lambda = 0.001$  to a near-random accuracy of 53% when  $\lambda = 1$ ).

In terms of unconditional gaps, adopting the adversarial strategy - at least for these levels of penalization - does not reduce the gender gaps to zero for all characteristics, as would perhaps have been expected. Indeed, statistically significant differences in terms of contract type, occupations, hours worked and fit to search criteria remain. Yet, for all values of  $\lambda$ , all unconditional gender gaps are considerably reduced. For instance, the log wage gap is divided by 12 (when comparing  $\lambda = 0$  to  $\lambda = 1$ ). All conditional gender gaps are also decreased.

Altogether, the use of adversarial de-biasing techniques, aiming at making recommendation gender-blind, entails a slight loss in recommendation performance. Moreover, it reduces unconditional and conditional gender gaps, without suppressing them.

Note that the presented adversarial strategy decorrelates the latent from gender, regardless of whether it represents features from job search fundamentals  $Z$  or from  $X \setminus Z$ . An adaptation of the strategy aiming to only target gaps conditional on  $Z$  is left for further work.

	$\lambda = 0$	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 are presented on the subsample of hired job seekers, for different weights  $\lambda$  given to the adversarial term in the loss function. Column  $\lambda = 0$  restates the standard algorithm's performances for convenience in comparisons. Recall and adversary accuracy are computed on the test set (all hired job seekers). Unconditional and conditional gaps are computed on the population of hired job seekers with common support. Unconditional gaps correspond to a difference in means between men and women. Conditional gaps are obtained by DML, using random forests to estimate  $m$  and  $p$ .

**Table 3**  
**Adversarial correction: recall, unconditional and conditional gender gaps**

## 6. Conclusion / perspectives

Our main contribution is an audit of the gender fairness of the MUSE recommender system, trained on real-world hiring data. First, we find recall to be slightly higher for women than for men. Second, we provide evidence of differentiated treatment of men and women by the algorithm in terms of recommended job characteristics, even conditionally on job seekers' search criteria. In the latter case, we find female job seekers to be recommended jobs that fit their own search criteria less often. In the latter case, we find female job seekers to be recommended jobs that do not increase gendered gaps observed in hirings or applications, and even decreases them in the cases of occupation type and working hours. A comparison of recommended job ads to application behavior leads to similar conclusions. Finally, we investigate the trade-offs between recommendation performance and gender gaps entailed by the use of adversarial de-biasing techniques. The use of such techniques entails a slight loss in terms of recall, but narrows some of the conditional and unconditional gender gaps without eliminating them.

Ultimately, the merits of de-biased algorithms attempting to reduce gender gaps in recommendations hinge on the acceptability of the proposed job ads in terms of job seekers' (possibly gendered) preferences. An algorithm straying off too far from job seekers' search behavior might lead to a deadweight loss: a loss in recommendation quality without any effect on labor market inequalities if recommendations are simply discarded as irrelevant. Answering whether a suitable equilibrium can be found requires interacting with job seekers.

## Acknowledgments

We warmly thank C. Vessereau, S. Robidou and P. Beurnier from *Pôle emploi* for making this research possible and granting access to the proprietary data. First author was funded on a grant from the DataIA Institute, Saclay.

## References

- [1] M. Belot, P. Kircher, P. Muller, Providing advice to jobseekers at low cost: An experimental study on online advice, *The review of economic studies* 86 (2019) 1411–1447.
- [2] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning. [fairmlbook.org](http://fairmlbook.org), 2019.
- [3] G. Bied, S. Nathan, E. Perennes, M. Hoffmann, P. Caillou, B. Crépon, C. Gaillac, M. Sebag, Toward job recommendation for all, Working paper (2023).
- [4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.
- [5] H. Varian, Efficiency, equity and envy, *Journal of Economic Theory* 9 (1974) 63–91.
- [6] M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, A. Weller, From parity to preference-based notions of fairness in classification, *Advances in neural information processing systems* 30 (2017).
- [7] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, Double/debiased/neyman machine learning of treatment effects, *American Economic Review* 107 (2017) 261–265.
- [8] H. Edwards, A. Storkey, Censoring representations with an adversary, 2015. URL: <https://arxiv.org/abs/1511.05897>. doi:10.48550/ARXIV.1511.05897.
- [9] P. Bach, V. Chernozhukov, M. Spindler, Closing the us gender wage gap requires understanding its heterogeneity, *arXiv preprint arXiv:1812.04345* (2018).
- [10] X. Nie, S. Wager, Quasi-oracle estimation of heterogeneous treatment effects, *Biometrika* 108 (2021) 299–319.
- [11] M. D. Ekstrand, A. Das, R. Burke, F. Diaz, et al., Fairness in information access systems, *Foundations and Trends® in Information Retrieval* 16 (2022) 1–177.
- [12] Y. Wang, W. Ma, M. Zhang\*, Y. Liu, S. Ma, A survey on the fairness of recommender systems, *ACM Journal of the ACM (JACM)* (2022).
- [13] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, Y. Zhang, Fairness in recommendation: A survey, 2022. URL: <https://arxiv.org/abs/2205.13619>. doi:10.48550/ARXIV.2205.13619.
- [14] A. Singh, T. Joachims, Fairness of exposure in rankings, in: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18*, Association for Computing Machinery, New York, NY, USA, 2018, p. 2219–2228. URL: <https://doi.org/10.1145/3219819.3220088>. doi:10.1145/3219819.3220088.
- [15] A. Singh, T. Joachims, Policy learning for fairness in ranking, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 32, Curran Associates, Inc., 2019. URL: <https://proceedings.neurips.cc/paper/2019/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf>.

- [16] S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2019. URL: <https://doi.org/10.1145/2F3292500.3330691>. doi:10.1145/3292500.3330691.
- [17] R. Mehrotra, A. Anderson, F. Diaz, A. Sharma, H. Wallach, E. Yilmaz, Auditing search engines for differential satisfaction across demographics, in: Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion, ACM Press, 2017. URL: <https://doi.org/10.1145/2F3041021.3054197>. doi:10.1145/3041021.3054197.
- [18] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, M. S. Pera, All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 172–186.
- [19] A. B. Melchiorre, N. Rekabsaz, E. Parada-Cabaleiro, S. Brandl, O. Lesota, M. Schedl, Investigating gender fairness of recommendation algorithms in the music domain, Information Processing & Management 58 (2021) 102666.
- [20] R. Islam, K. N. Keya, Z. Zeng, S. Pan, J. Foulds, Debiasing career recommendations with neural fair collaborative filtering, in: Proceedings of the Web Conference 2021, 2021, pp. 3779–3790.
- [21] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Enhancement of the neutrality in recommendation, in: Decisions@ RecSys, 2012, pp. 8–14.
- [22] C. Rus, J. Luppès, H. Oosterhuis, G. H. Schoenmacker, Closing the gender wage gap: Adversarial fairness in job recommendation (2022). URL: <https://arxiv.org/abs/2209.09592>. doi:10.48550/ARXIV.2209.09592.
- [23] S. Zhang, P. Kuhn, Understanding algorithmic bias in job recommender systems: An audit study approach (2022).
- [24] C. Wadsworth, F. Vera, C. Piech, Achieving fairness through adversarial learning: an application to recidivism prediction, 2018. URL: <https://arxiv.org/abs/1807.00199>. doi:10.48550/ARXIV.1807.00199.
- [25] A. Beutel, J. Chen, Z. Zhao, E. H. Chi, Data decisions and theoretical implications when adversarially learning fair representations, 2017. URL: <https://arxiv.org/abs/1707.00075>. doi:10.48550/ARXIV.1707.00075.
- [26] Y. Li, H. Chen, S. Xu, Y. Ge, Y. Zhang, Towards personalized fairness based on causal notion, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2021. URL: <https://doi.org/10.1145/2F3404835.3462966>. doi:10.1145/3404835.3462966.
- [27] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American society for information science 41 (1990) 391–407.
- [28] G. W. Imbens, D. B. Rubin, Causal inference in statistics, social, and biomedical sciences, Cambridge University Press, 2015.
- [29] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
- [30] T. Le Barbanchon, R. Rathelot, A. Roulet, Gender differences in job search: Trading off commute against wage, The Quarterly Journal of Economics 136 (2021) 381–426.

- [31] N. Fortin, T. Lemieux, S. Firpo, Decomposition methods in economics, in: Handbook of labor economics, volume 4, Elsevier, 2011, pp. 1–102.
- [32] M. P. Kim, A. Korolova, G. N. Rothblum, G. Yona, Preference-informed fairness, arXiv preprint arXiv:1904.01793 (2019).
- [33] P. M. Robinson, Root-n-consistent semiparametric regression, *Econometrica: Journal of the Econometric Society* (1988) 931–954.

## A. Additional tables

	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

Notes: The first column presents the total sample size for the different datasets used in the analysis: "Full week" and "Full week (overlap)" present the sample size for a week in the test set before and after restriction to job seekers satisfying the overlap condition required in the Double Machine Learning method of Section 3, "Hires", "Hires (overlap)", and "Hires & Applications (overlap)" present respectively the sample sizes for the subsamples of job seekers in the test set who have been hired, hired and for whom the overlap condition holds, and the subset of the latter one where we also observe applications.

**Table 4**  
Size of datasets used for the analysis

Top $k$	Recall@ $k$	Men	Women	p-value
1	0.0555	0.0528	0.0578	0.021
5	0.1751	0.1655	0.1835	0.000
10	0.2557	0.2432	0.2666	0.000
20	0.3508	0.3333	0.3660	0.000
50	0.4905	0.4747	0.5044	0.000
100	0.5902	0.5761	0.6026	0.000

Notes: Recall@ $k$  is the recall on all the population on the first top  $k$  recommendations. Columns "Men" and "Women" present the same recall@ $k$  separately for men and women. The last column performs a test of equality between columns 2 and 3.

**Table 5**  
Difference in recall between genders.



## B. Details on variables used

Table 6 presents the comprehensive list of variables included in vector  $Z$  describing jobseekers job search parameters and qualifications. All remaining variables related to jobseekers characteristics and included in the algorithm are given in Table 7.

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)

**Table 6**

Jobseekers related features included in vector  $Z$

Socio-demographic variables	
Number of children	numeric
Jobseeker lives in a QPV area <sup>4</sup>	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

**Table 7**

Jobseekers related features given to the algorithm for the generation of recommendations but not considered in the vector  $Z$

<sup>4</sup>QPV refers to poor urban areas in need of public intervention, particularly in terms of urban renewal

### C. Details on the estimation of the heterogeneous effect of gender on the recommendations using the double machine learning method (DML).

To perform the estimation of the *gender gap*  $\tau$ , we use the double machine learning method (DML) [see, e.g., 7, 10]. This method is based on a rewriting of (1), following the intuition of [33], as

$$Y - m(Z) = (G - p(Z))\tau + \varepsilon, \quad \mathbb{E}(\varepsilon|Z, G) = 0, \quad (2)$$

where  $m(Z) := \mathbb{E}(Y|Z)$  is a regression function and  $p(Z) := \mathbb{P}(G = 1|Z)$  is the propensity score, i.e. the probability to be a women ( $G = 1$ ) given the observed preferences and qualifications  $Z$ . The later are *nuisance parameters*, which have to be estimated in a first step, but the reformulation (3) allows the estimation of  $\tau$  to be *doubly robust* to this first stage estimation error. This means that we can obtain an estimator for  $\tau$  which is asymptotically normal under theoretical conditions which are satisfied by many machine learning methods. Estimation thus consists of 1) estimating  $m$  and  $p$  using machine learning estimators  $\hat{m}$  and  $\hat{p}$ ; 2) estimate the gender gap  $\tau$  via minimization of the mean squared error associated to (3) using plug-in leave-one-out versions of  $\hat{m}$  and  $\hat{p}$ , i.e., i.e. predicting without using the  $i$ -th example [see 10].

## D. Gendered recommendations, applications, and hires: a simple formal model

To discuss the different sources of biases which can appear in the recommendations, and how they compare to those appearing both in the realized job applications and hires, we consider the following simple model of the decision to apply for a job and of the hiring.

For job seekers and job ads having respective types  $x$  and  $z$ , we denote the chances that the interview yields a hiring by  $\pi(x, z)$ . The job seekers may not be rational and have expectations about their opportunities  $\tilde{\pi}(x, z)$  which differ from the objective ones  $\pi(x, z) \neq \tilde{\pi}(x, z)$ . We assume that job seekers expect to have a utility  $U(x, z) + \varepsilon - k$  if hired, where  $\varepsilon$  is a unobserved random part and  $k$  is the cost of application. On the contrary, they expect to have their baseline utility  $U_0(x)$  minus the cost  $k$ . In this model, job seekers decide to apply for a job with type  $z$  if their expected utility if they do so is greater than their utility if they do not apply, namely  $U_0(x)$ :

$$\text{(Decision applying)} \quad \underbrace{\tilde{\pi}(x, z)(U(x, z) + \varepsilon) + (1 - \tilde{\pi}(x, z))U_0(x) - k}_{\text{Expected utility when applying}} \geq \underbrace{U_0(x)}_{\text{Utility without applying}} .$$

In this model, the probability of observing an application of  $x$  on a job ad of type  $z$  is

$$\text{(Probability of observing an application)} \quad A(x, z) = F_{-\varepsilon} \left( U(x, z) - U_0(x) - \frac{k}{\tilde{\pi}(x, z)} \right),$$

where  $F_{-\varepsilon}$  denotes the cdf of  $-\varepsilon$ . We note that, when the cost of application are zero,  $k = 0$ , we find the intuitive idea that only the utility matters in the job seekers' decisions. Otherwise, their expected chances  $\tilde{\pi}(x, z)$  of a positive output weight their utility gains and might censor their decision of applying, hence the observed data. This simply underlines that realized applications are then not a pure expressions of the preferences, but also mix with possibly wrong expectations. It is finally of interest to consider the form taken by the probability of observing a hiring, which is simply the product of the probability of application times the objective probability of a positive output after the interview:  $H(x, z) = A(x, z)\pi(x, z)$ .

This model helps us discussing several mechanisms that could yield a differential treatment along the lines of gender. First, preferences  $U$  might be gender-specific [30], *e.g.* women tend to appreciate the relative values of commuting time and wages differently from men. An algorithm learning from past hires or applications could reproduce these differences in preferences. If the later are the product of social norms or other constraints, a policymaker might find this unfair that the algorithm convey these differences, hence justifying to impose parity along these lines. Second, even if job seekers are rational, there might be gendered differences in the hiring chances  $\pi$ , *e.g.* taste or statistical discrimination against a gender by recruiters.

The algorithm could also reproduce these differences. A final pitfall is that the hiring expectations  $\tilde{\pi}$  might also be gendered: there might be differences in the perceptions and the representations of the chances to be hired, leading to differences in self-censorship or over-confidence. In our model, this could directly create or exacerbate the differences which might already be present in the objective chances  $\pi$ , and impacting the training data of the algorithm.

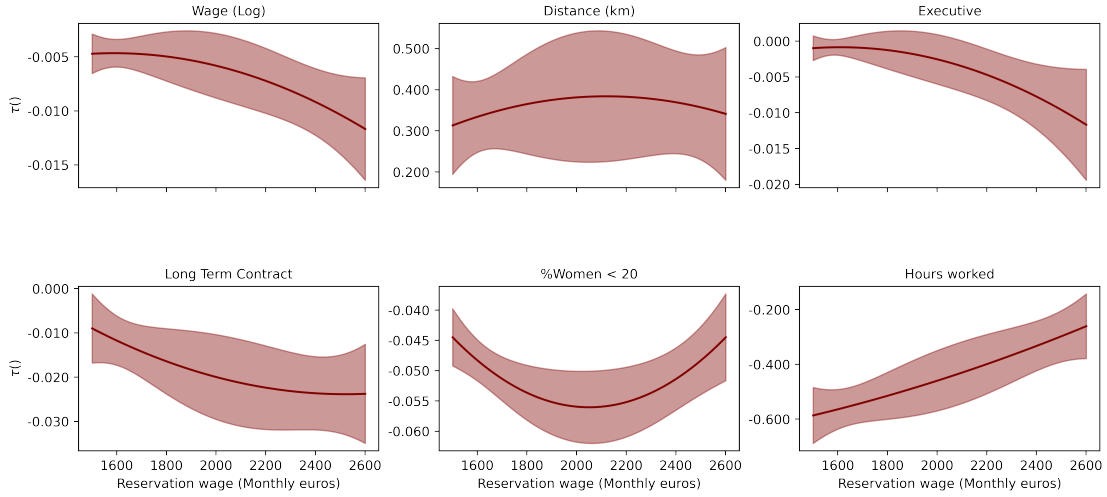
## E. Robustness checks

To ensure estimates for  $\tau$  obtained by Double Machine Learning results are robust to the choice of machine learning technique used for the approximation of  $m$  and  $p$ , we report alternative estimates for  $\tau$  obtained using a XGBoost and Lasso estimators as well as the p-values associated in Table 8, columns 3-6. Results are consistent with what we find with a random forest estimator (Columns 1-2).

	Cond. $\tau$ <i>Random Forest</i>	p-value	Cond. $\tau$ <i>XGBoost</i>	p-value	Cond. $\tau$ <i>Lasso</i>	p-value
Wage (log)	-0.004	0.000	-0.010	0.000	-0.006	0.000
Distance (km)	0.400	0.000	0.145	0.000	0.331	0.000
Executive	-0.002	0.032	-0.005	0.000	-0.002	0.032
Long term contract	-0.014	0.000	-0.025	0.000	-0.012	0.000
%Women < 20	-0.033	0.000	-0.110	0.000	-0.035	0.000
Hours worked per week	-0.381	0.000	-1.025	0.000	-0.463	0.000
Fit to job search parameters	-0.011	0.000	-0.014	0.000	-0.008	0.000

Notes: Column 1, 3 and 5 report, on the population of job seekers with sufficiently comparable characteristics, the estimates for the gender gap  $\tau$  controlling for search parameters using DML and respectively a random forest, XGBoost and lasso estimator for the functions  $m$  and  $p$ .

**Table 8**  
**Conditional gender differences in characteristics of offers recommended using different estimation methods**



**Figure 1:** Gender gaps according to Reservation Wage

## F. Heterogeneity

Our main specification (equation 1) focuses on average gender gaps  $\tau$  (after controlling for job search fundamentals  $Z$ ). However, gender gaps are likely to be heterogeneous, at least for a subset  $Z_0$  of  $Z$ . For instance, the gender gaps in recommendations may be greater for women looking for high wages than for those seeking low wages. Accordingly, we propose to study gender gaps conditional on  $Z_0$ , in line with the estimation of so-called Conditional Average Treatment Effects in the causal estimation literature [28]. More precisely, to provide insights about this potential heterogeneity, we assume

$$Y - m(Z) = (G - p(Z))\tau(Z_0) + \varepsilon, \quad \mathbb{E}(\varepsilon|Z, G) = 0, \quad (3)$$

with  $\tau(Z_0)$  a linear function. In the following, we consider  $Z_0$  as an expansion of a single feature of interest - job seekers' monthly reservation wage in euros - on a base of B-splines to increase the specification's flexibility. To reduce sensitivity on outliers we top code at 90% and bottom code at 10%.

Figure 1 shows the conditional gender wage gap (solid line) in the characteristics of recommendations according to reservation wage and provides confidence interval at 95%.