Fairness in Algorithmic Profiling: The AMAS Case

EVA ACHTERHOLD, LMU Munich, Germany
MONIKA MÜHLBÖCK, University of Vienna, Austria
NADIA STEIBER, University of Vienna, Austria
CHRISTOPH KERN, LMU Munich, Germany

We study a controversial application of algorithmic profiling in the public sector, the Austrian AMAS system. AMAS was supposed to help case workers at the Public Employment Service (PES) Austria to allocate support measures to job seekers based on their predicted chance of (re-)integration into the labor market. Shortly after its release, AMAS was criticized for its apparent unequal treatment of job seekers based on gender and citizenship. We systematically investigate the AMAS model using a novel real-world dataset of young job seekers from Vienna, which allows us to provide the first empirical evaluation of the AMAS model with a focus on fairness measures. We further apply bias mitigation strategies to study their effectiveness in our real-world setting. Our findings indicate that the prediction performance of the AMAS model is insufficient for use in practice, as more than 30% of job seekers would be misclassified in our use case. Further, our results confirm that the original model is biased with respect to gender as it tends to (incorrectly) assign women to the group with high chances of re-employment, which is not prioritized in the PES' allocation of support measures. However, most bias mitigation strategies were able to improve fairness without compromising performance and thus could form an important building block in the development of fair profiling systems in the present context.

CCS Concepts: • **General and reference** → *Metrics*; *Evaluation*; *Metrics*; *Evaluation.*

Additional Key Words and Phrases: algorithmic profiling, statistical discrimination, public employment services, artificial intelligence, bias mitigation

ACM Reference Format:
Eva Achterhold, Monika Mühlböck, Nadia Steiber, and Christoph Kern. 2023. Fairness in Algorithmic Profiling: The AMAS Case. 1, 1 (September 2023), 23 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Algorithmic profiling is increasingly used in high-stake decision making where incorrect predictions can have a profound impact on an individual's life. Data-driven decision making systems are being used in areas such as justice [20], healthcare [42] and credit scoring [28]. In a world that is becoming more and more complex, algorithmic profiling systems enable the integration of vast amounts of data and thus promise to be more reliable [6], efficient [38], transparent [52], and accountable [33] than human decision-making. Further, relying on statistical models for consequential decision-making offers the presumed advantage that the results do not depend on individual decision makers and are therefore more objective and consistent. Previous research has shown that data-driven methods are in fact able to outperform humans in terms of accuracy in prediction tasks [50].

The promise of decision neutrality through the use of algorithms, however, has been refuted many times. One of the most prominent examples of discrimination through algorithmic profiling is COMPAS, an algorithm that predicts a defendant's recidivism risk to help judges decide whether to detain or release the defendant. When comparing error rates

Authors' addresses: Eva Achterhold, mail@acheva.de, LMU Munich, Munich, Germany; Monika Mühlböck, University of Vienna, Vienna, Austria; Nadia Steiber, University of Vienna, Vienna, Austria; Christoph Kern, LMU Munich, Munich, Germany.

2023. Manuscript submitted to ACM

between black and white defendants, it was found that black defendants were more likely to be misclassified as future offenders, while white defendants were more often incorrectly classified as low-risk [5]. In addition, differences between subgroups could not be explained by prior crimes, future recidivism, age, or gender. Thus, the attribute *race* played a crucial role in the decision-making process. This is not only problematic in that it contradicts anti-discrimination legislation, but it also undermines efforts to overcome biases that exist in society.

The controversy about the COMPAS system has been mainly ignited in the United States, but there has also been a recent debate in Europe about the discriminatory side effects of algorithmic profiling. In 2018, the Public Employment Service (PES) Austria (AMS, Arbeitsmarktservice) introduced AMAS (Arbeitsmarkt-Chancen Assistenzsystem [Labor market opportunities assistance system]), a system that was intended to support case workers with the decision of allocating service resources to job seekers, in a pilot phase [25]. The idea was to supplement the case worker's subjective assessment with an standardized data-driven evaluation of a person's chances of re-employment. The expected benefits of this added process automation were twofold: First, it should increase the effectiveness of labor market programs by targeting support measures to individuals who will benefit most from them. Second, it should improve the effectiveness as well as the efficiency of the process, meaning that case workers should provide the most accurate assessment of the need for assistance in the shortest time possible in order to process more cases and allocate resources optimally [2].

Upon registration with the AMS, a so-called Integration Chance (IC) score was calculated for each individual based on their labor market history and personal characteristics. To account for potentially incomplete data (e.g., from immigrants) or fragmented employment histories (e.g., from young adults), specific model variants were developed for four groups of job seekers. For a differentiated analysis, the criterion re-employment was defined in two different ways: In the short-term perspective, individuals who were employed for at least 90 days within the seven months after reporting unemployment were counted as having high prospects of employment. In the long-term perspective, the threshold was raised to 180 days of employment within 24 months.

Based on the calculated IC score, job seekers were placed into one of three categories: Group A consists of individuals with a greater than 66% probability of short-term re-employment. Since the assumption is that these persons are not difficult to integrate, fewer measures are to be assigned to this group. Group C, on the other hand, consists of those individuals for whom the model predicts less than a 25% IC within the long-term criterion. This group is passed on to external service providers for efficiency reasons, but will not receive support measures from the AMS. All other persons are assigned to group B and thus fall under the target group of AMS labor market measures.

Shortly after publishing the model, criticism was raised by several researchers [2, 12, 40], journalists [48] and privacy groups [14]. One of the main issues was the lack of transparency of the algorithm. At first, neither the data on which the calculations were based nor a detailed description of the model itself was provided. Upon request, a paper describing a logistic regression model was shared including regression coefficients that can be used to predict the risk score for short-term unemployment [25]. Although, as clarified later, the logistic regression model is supposed to serve as a representation of a stratification procedure actually used, the documentation revealed that the two attributes gender and citizenship had a negative impact on predicted re-employment chances [25]. This implies that according to the main AMAS model women, as well as individuals with non-EU citizenship, are less likely to be integrated into the labor market in the short-term (three months employment in seven months after registration). Thus, the system was criticized for reflecting historical discrimination in the labor market with respect to gender [9] and ethnicity [53]. In 2020, the Austrian Data Protection Authority (DPA) prohibited AMAS, arguing that a legal basis for conducting "profiling" was missing [31]. The AMS appealed against the notice and won at the Federal Administrative Court, with the DPA appealing against the court's decision. The case is currently pending before the Austrian Supreme Administrative Court.

The outlined concerns regarding possibly biased predictions and discriminatory allocation processes paired with the scope and potential impact of the system highlight the need for a systematic fairness audit of the AMAS approach. While Allhutter et al. [2] study the AMAS by means of a document analysis, an empirical investigation with a focus on fairness measures and bias mitigation techniques is lacking. Given a novel real-world dataset of young job seekers from Austria, our work, to our knowledge, provides the first empirical case study of the AMAS system.

Next to our fairness evaluations, we assess the ability of de-biasing techniques in mitigating potential biases of the AMAS model. While many studies on bias mitigation techniques focus on prediction tasks from a small set of benchmark data [18, 44], we set out to compare de-biasing techniques in the labor market context based on an existing profiling approach.

We contribute to the growing topic of fairness in algorithmic profiling in the public sector by studying a high-stake profiling model using real-world data of job seekers that fall into the models' "target group". The main findings of our study are as follows:

- The prediction performance of the AMAS model on our data set is mediocre at best and leaves about 30% of job seekers misclassified.
- We observe considerable differences in statistical parity, true positive and false positive rates between male and female job seekers based on the AMAS model predictions.
- Bias mitigation strategies are able to reduce bias in the model results and are accompanied by a tolerable drop in performance. Although the choice of the classification threshold affects the performance and fairness metrics, we find that de-biasing methods are effective over various thresholds.

2 BACKGROUND AND RELATED WORK

2.1 Fairness considerations in unemployment profiling

The use of algorithms to support the allocation of limited public resources has become increasingly common in recent years. In many countries, PESs use algorithmic profiling tools to identify individuals who are at high risk of Long-Term Unemployment (LTU) [41]. To prevent LTU, support measures are assigned to a selected group of job seekers with a similar predicted risk. Across countries, the systems differ in terms of the predictors used for classification (e.g., administrative records, questionnaires), the prediction criterion (e.g., LTU, re-employment chances), the classification model used (e.g., logistic regression, random forest), and the allocation strategy (e.g., supporting those at highest risk for long-term unemployment, identifying the optimal treatment for an individual). For a comprehensive overview of different approaches, we refer the interested reader to Loxha et al. [41] and Desiere et al. [15].

Given the example from Austria, fairness implications for allocation of scarce resources arise. The distribution of benefits and burdens has long been a topic in philosophy and is closely linked to debates about equality, equity, and justice. The discourse on distributive justice, i.e., the consideration of just allocation of resources among members of a society [37], has resulted in a variety of theories with different approaches to ensuring fairness.

Recent efforts have been made to integrate the philosophical perspective with mathematical formalism of fairness metrics for Fair Machine Learning (FairML) [7, 35]. However, several aspects of fairness originate from normative determinations in society that cannot be accounted for by mathematical approaches. In any decision scenario, an action is taken based on a decision rule that is more or less strictly specified. Thus, the question of distributive justice arises even in human decision-making. However, in data-driven decision-making, there is the additional question of

a fair prediction to which the decision rule is then applied. Kuppler et al. [34] propose to distinguish between *fair predictions* as an aspect related to algorithmic output and *just decisions* related to the outcome of the decision made. This differentiation can be applied to the context of algorithmic profiling of the unemployed to illustrate the impact of technical fairness-enhancing interventions on the socio-technical decision-making process of allocating public resources.

Just decisions. The question of just decisions in the allocation of public resources is not limited to the use of algorithmic systems. It is still common in many countries that allocation of PES support measures is done by case workers, either by relying solely on their expertise or by following rules such as passing a threshold for time in unemployment [41]. As already stated, these human decision-making processes are not free from bias. To quantify justice in decisions, we may study how actions are allocated to social groups. Thus, measures that fall under the *independency criterion* (e.g., Statistical Parity Difference, Disparate Impact) [6] can be used to evaluate to what extend social groups are treated differently in the decision-making process.

In their study on justice in decisions, Kern et al. [27] applied different classification methods to predict the risk of becoming LTU on German data. They could show that although the statistical models presented had a similar level of accuracy they have very different fairness implications. Specifically, they found that the models tended to reinforce parity differences between individuals belonging to an unprivileged group (female, non-German) compared to the privileged group (male, German).

Under the criterion of independence, we study whether the chance of receiving support resources depends on the gender of a person. As a measure, we use Statistical Parity Difference and an adapted version of Disparate Impact (see subsubsection 3.3.2). However, the assumption of having the same right to receiving the public resource implies that societal groups have similar (true) chances of reintegration into the labor market. This is countered by the fact that studies have found structurally different integration opportunities, for example for women [4, 45]. An algorithm for guiding the distribution of support measures that meets the independency criterion would not sufficiently account for these differences.

Fair predictions. What the use of algorithmic profiling to allocate public resources adds to the debate on fairness is the step of deriving the decision from a prediction made by an algorithm. Even if, in a fictitious world, we can ensure a just decision rule, predictions that are biased could still encode unequal treatment in the decision process. Thus, in order to assess the fairness of predictions, we need to take into account both the observed and the predicted outcome, which is done with the *separation criterion* and corresponding fairness objectives (e.g, Equal Opportunity, Equalized Odds) [6].

The separation criterion requires that the error rates of the classifier are equal across groups, under the assumption that the given distribution of the actual outcome is representative for it. In the context of the distribution of public resources, this means that, assuming someone is actually not re-employed, the model should predict a poor chance of integration for that person. However, if the reintegration chance given by the true label differs between groups, this should also be reflected in the prediction.

Several researchers have investigated the data-driven allocation of public resources with respect to fair predictions, although with few studies focusing on profiling models for the unemployed [32]. Desiere and Struyven [16] investigated fairness implications of an algorithmic profiling tool that is used by the Flemish PES *VDAB* in Belgium. They found that the classifier was more likely to predict a high risk of LTU for job seekers belonging to a historically disadvantaged group, such as people of non-Belgian origin, people with disabilities, or the elderly. This inequality, measured as the ratio of False Positive Rate (FPR) between groups, was more prevalent in the predictions of the algorithmic profiling approach than in a simple rule-based approach, although accuracy was higher for the former. In addition, the authors

show that the bias depends strongly on the threshold used to distinguish the high-risk group from the low-risk group, as the proportion of minority groups decreased at higher thresholds.

The aforementioned studies examining the fairness of algorithmic profiling systems for the unemployed [16, 27] are among the few empirical research efforts in this area. Although both studies have shown that systems tend to produce biased outcomes and highlight that the decision to receive support from the PES can have a significant impact on an individual's life, this topic has received limited attention in fairness research to date. One possible reason is the difficulty in obtaining detailed data on job seekers' (un)employment histories and the lack of access to the actual systems.

With respect to the AMAS use case, Allhutter et al. [2] provide a systematic document review emphasizing the socio-technical implications and consequences of the use of the proposed system. Lopez [40] extends the analysis of the AMAS system by also raising issues of intersectional discrimination, legislation, and the efficiency of the system. Both works, however, do not provide empirical analysis or a data-based fairness assessment of the AMAS system.

Our work contributes to the limited body of literature on fairness of unemployment profiling systems by combining the practical application of various fairness metrics and bias mitigation strategies to a real-world use case. Before turning to the methods used in this study, we will briefly shed light on the implications of misclassification in the context of allocating support measures.

2.2 Implications of misclassification

The use of an algorithmic profiling system to support the allocation of public resources typically follows the aim of efficiently distributing the PES measures, i.e., providing support to those job seekers who actually need it. To assess this *need*, a model is trained to predict either the risk of LTU or a job seeker's chances of re-integration. A threshold t is then set to determine on the basis of the prediction whether someone is classified in the group of those who will be supported or those who will not. Note that here we simplify the mapping from the actual re-employment outcome, which e.g. includes information about whether a person was employed for at least 90 days within seven months of registration, to the conclusion that someone who is re-employed does not need support measures from the PES. In reality, there are many other factors to consider in this mapping, but their inclusion is beyond the scope of this work.

Given the binary classification for an instance and the information about the actual outcome that we know from evaluation data, we can evaluate which instances have been misclassified, i.e., assigned a different predicted outcome than the actual one. Misclassification can occur in two cases: First, if the algorithm wrongly predicts a negative outcome, which is referred to as a False Negative (FN), and second if the prediction is positive given the actual outcome is negative, which is referred to as a False Positive (FP). What is important to note at this point is that misclassification is always costly and changes with variations of t.

However, the cost of misclassification differs depending on the perspective under which the algorithm is evaluated. To assess the performance of a classification model used for allocating support measures, we need to take into account the resulting social implications. We have added Table B.1 to illustrate types of errors and the resulting consequences. Considering the PES objective of cost-efficient allocation, any person who does not need measures but still receives them imposes additional costs. This is the case when the algorithm predicts a negative outcome (no re-employment predicted - receives measures) when in fact there is a positive outcome, i.e., FN. Thus, from a PES perspective, a good algorithm is one where the False Negative Rate (FNR) is low, i.e., where among all actual positive outcomes, only a few are incorrectly predicted to be negative. Since $FNR = 1 - TPR = 1 - Recall$, a good classifier for the PES will have high recall and thus low FNR.

If we consider the job seekers' perspective, however, the greater disadvantage is not for those who receive measures without justification (FN), but for those who do not receive measures even though they would have needed them (FP). As a measure that takes FP into account, we use precision (see subsubsection 3.3.1), which tells us the proportion of individuals who are assigned a positive outcome (re-employed predicted - no measures) and were actually successfully reintegrated. It follows that if we take the False Discovery Rate (FDR) defined as $FDR = 1 - Precision$ we get the proportion of positively predicted individuals who were misclassified and would have actually needed support. From the job seekers' perspective, we therefore aim for high precision so that the FDR becomes low. Overall, we thus aim for high recall and precision and consequently high F1 values for all models studied.

In accordance with the AMAS model, we set the classification threshold for our performance and fairness assessment to $t = 0.66$ [25]. This choice actually leads to higher precision at the cost of recall than with the usual threshold of 0.5, as can be seen in Figure A.1. Thus, the initial AMS decision tends to be in favor of the job seeker, as the fraction of individuals who are FP tends to decrease with increasing thresholds.

3 METHODS

3.1 Data

The data used for this study was obtained as part of the panel survey "JuSAW – Jung und auf der Suche nach Arbeit in Wien" [Young and looking for work in Vienna] [46, 47]. Between April and September 2014, a total of 1246 individuals between the age of 18 and 28 who registered with the AMS as job seekers in Vienna participated in the study.

The aim of the study was to investigate the causal effects of unemployment on factors such as psychological and mental health, attitudes, and values. For this purpose, a first interview was conducted shortly after entering the registered job search and a second one a year later. The resulting data was linked to administrative records on employment history, education and socio-demographic attributes of the participants.

For our study, we selected individuals that are younger than 25 as this subgroup represents the target population that would be assessed by the AMAS model for young adults. Further, we extracted a set of 15 variables from the JuSAW data that match the ones used for AMAS [22, 25]. For each individual in our dataset, a binary variable indicates the `re-employment` outcome of finding a job with respect to the short-term criterion as defined by AMAS (three months of employment in seven months after registration). According to the AMAS documentation, young adults would not be assessed under the long-term criterion as assignment to external supervision is inadmissible for this group [3, 49]. We are thus left with a classification into group A (high re-employment chance - no measures) or B (lower chances - receives measures). After processing the data, our dataset includes $n = 678$ young job seekers and is split into 70% for model training and de-biasing and 30% for evaluation.

At this point, it is important to note that almost 28% of the job seekers in our dataset entered an AMS-funded support measure between the first and second interview, which in the original study had a significant negative effect on the outcome variable `re-employment` [47]. This indicates the so-called "lock-in" effect, according to which participation in active labor market programs increases the time in unemployment due to reduced time resources for job search or awaiting a desired qualification. However, the influence of interventions is also taken into account in the AMAS model, as the original model uses data on past support measures as input (see Table B.2). More specifically, in the initial AMS data, individuals who had already participated in measures were also included in the assessment of the re-employment rate under the short-term criterion with the aim of improving the future allocation of resources.

We closely resemble the AMAS process in our encoding of the predictor variables. Similar to the AMS, we included `health impairment`, which was originally measured with a four point scale in the JuSAW questionnaire, as a binary feature in our analysis. The original AMAS model further takes into account `obligations of care` only for women. As there is no detailed information in the AMAS documentation on how obligations of care are defined, we used the data from the questionnaire on the number of children of a person. To make our model as close to the AMAS algorithm, we removed information on `obligations of care` for male individuals in our dataset.

3.2 Prediction

As mentioned earlier, the procedure for determining an individual's IC score with AMAS is based on a stratification analysis of a dataset of job seekers in Austria. While the exact procedure has not been disclosed publicly, the AMS has published coefficients of a Logistic Regression (LR) that allow to construct a simplified model of the system [22, 25].

3.2.1 Prediction setup. The prediction task is based on the following components:

- Set of **predictor variables** X which include sensitive and nonsensitive attributes (see Table B.2).
- **Protected attribute** $A \in \{p, u\}$, where $A = p$ indicates members of the protected group and $A = u$ those of the unprotected group.
- **Observed outcome** $Y \in \{0, 1\}$. Instances that were re-employed for at least 90 days within seven months after registration are labelled as $Y = 1$ and those that were not re-employed as $Y = 0$, respectively.
- **Risk score** $R \in [0, 1]$ that is equivalent to the IC score.
- **Prediction** $\hat{Y} \in \{0, 1\}$. Binary prediction that is obtained by setting $\hat{Y} = 1\{R > t\}$ where 1 denotes the indicator function and t is a threshold to be set.

3.2.2 Prediction models. We use two different methods for predicting re-employment chances of job seekers:

- **AMAS.** An application of the original AMAS model for young adults using the coefficients published by [25] and shown in Table B.2.
- **LR.** Common logistic regression that is used as a benchmark. We set the class weight parameter to "balanced" to prevent predicting only the majority (negative) class and to reduce misclassification errors in the positive class.

3.2.3 Software. The analysis was carried out with *Python 3.9*. For data preparation, we used the *pandas* library. Model training and performance evaluation were done with the *scikit-learn* package. Fairness metrics and bias mitigation algorithms were provided by IBM's *AIF360 toolkit* [8].

3.3 Metrics

In order to evaluate the effectiveness of AMAS and bias mitigation algorithms, we compare the results with respect to prediction performance and fairness metrics.

3.3.1 Performance metrics. When it comes to classifying job seekers, one main aspect is to accurately distinguish those that are able to find a job without receiving resources from those that face high LTU risk. Bothmann et al. [10] even argue that predictive performance is a prerequisite for achieving fairness. All metrics listed below are used to evaluate performance based on predicted classes $\hat{Y}$ and can take values in range $[0, 1]$. As a distribution function we consider the empirical distribution measure P induced by the underlying dataset.

- **Accuracy.** $Acc = P(\hat{Y} = Y)$
- **Precision.** $Prec = P(Y = 1|\hat{Y} = 1)$
- **Recall.** $Rec = P(\hat{Y} = 1|Y = 1)$
- **F1 Score.** $F1 = 2 \times \frac{Prec \times Rec}{Prec + Rec}$
- **AUC.** Area Under Curve (AUC) as an aggregate measure of performance captures the two-dimensional area under the receiver operating characteristic curve, which plots the TPR against the FPR for multiple thresholds.

3.3.2 Fairness measures. We use four fairness metrics to assess the fairness of our classifiers following the discussion in section 2. All of the metrics use properties of the joint distribution of the sensitive attribute A, the true outcome Y and the binary prediction $\hat{Y}$. For all measures, a negative value indicates a bias of positive results in favor of the unprotected group, $A = u$. Consistent with studies showing that women tend to have lower chances of re-employment than men [4, 45], we decided to denote the group of male job seekers as unprotected ($A = u$) and that of female job seekers as protected ($A = p$), respectively.

- **Statistical Parity Difference (SPD) [17].** SPD measures the difference in positive outcomes between two subgroups that differ according to their protected attribute A. It is computed as follows:

$$SPD(\hat{Y}) = P(\hat{Y} = 1|A = p) - P(\hat{Y} = 1|A = u)$$

  Note that a SPD value can also be calculated for the actual outcome by replacing $\hat{Y}$ with Y.

- **Disparate Impact (DI) [19].** DI takes the ratio in positive prediction rates for both groups. This measure is formulated as follows:

$$DI(\hat{Y}) = \frac{P(\hat{Y} = 1|A = p)}{P(\hat{Y} = 1|A = u)}$$

  In order to interpret DI in the same way as the difference metrics, where group parity is indicated by a score of zero, we use a scaled DI measure, which we refer to as *Disparate Impact Scaled* (DIS) and define as follows:

$$DI_{scaled}(\hat{Y}) = \begin{cases} 1 - 1/DI(\hat{Y}), & \text{if } DI(\hat{Y}) > 1 \\ -1 + DI(\hat{Y}), & \text{if } DI(\hat{Y}) \leq 1 \end{cases}$$

  As with SPD, we can also evaluate DI and DIS for the actual outcome Y.

- **Equal Opportunity Difference (EOD) [24].** To quantify the disparity in true positives between groups based on the protected attribute, we calculate the following:

$$EOD = P(\hat{Y} = 1|A = p, Y = y) - P(\hat{Y} = 1|A = u, Y = y), \qquad y \in \{0, 1\},$$

  with $y = 1$. Besides focusing on the difference in TPR, we can also use the measure of EOD with respect to FPR by setting $y = 0$.

- **Average Odds Difference (AOD) [1].** A classifier's fairness with regard to Equalized Odds can be measured by the AOD, which is formulated as follows:

$$AOD = \frac{1}{2} \times \Big((P(\hat{Y} = 1|A = p, Y = 0) - P(\hat{Y} = 1|A = u, Y = 0)) + (P(\hat{Y} = 1|A = p, Y = 1) - P(\hat{Y} = 1|A = u, Y = 1)) \Big)$$

3.4 Bias mitigation

Our selection of bias mitigation techniques is restricted to algorithms that can deal with categorical input and a binary sensitive attribute. Since for this work we assume the existence of a classifier that has been introduced but was found to discriminate, such as in the AMAS example, we focus on the following pre- and post-processing techniques (see also Appendix C) and compare the original and mitigated results along the metrics presented in the previous section.

- **Reweighing (RW) [11].** The method of RW is a pre-processing technique that adjusts the weight of each example in the training data, based on its membership in different groups defined by the sensitive attribute. The method assigns different weights to different groups, with the goal of balancing the distribution of the protected attribute in the training data.

  The weight for each example in the training data is calculated as a fraction of the *expected probability*, which is calculated by multiplying the probability of being in a group by the probability of being in a particular class, and the *observed probability*, which is the actual probability of a certain group of individuals to be in a certain class, taking into account the sensitive attribute.

  Using the RW method on our data aims at reducing the dependency between predicted re-employment chances and the protected attribute `gender`. This implies the assumption that the result of a fair classifier should be independent of the protected attribute. Therefore, RW aims at balancing SPD and DI. Since the classifier needs to be retrained on the weighted training dataset (see Table B.4), we can only apply this method to the LR model but not to the AMAS model.

- **Learning Fair Representations (LFR) [51].** The LFR method aims to learn a new representation of data that is both predictive and fair by removing any bias present in the input data that could lead to biased predictions from a machine learning model. By generating a latent representation that retains all necessary information about an individual, but obfuscates group membership derived from a protected attribute, the aim is to ensure independence between the prediction and the sensitive attribute and thus balancing SPD and DI.

  The LFR method adds an additional constraint to the objective function that ensures the sensitive attribute cannot be inferred from the representation, i.e., minimizing the mutual information between the sensitive attribute and the learned representation. The resulting objective function is the sum of the reconstruction term, the fairness term, and the output prediction error. The trade-offs between these terms are governed by custom weights for the fairness constraint term A_z, the reconstruction term A_x, and the output prediction error A_y. These weights, as well as the number of prototypes k, are hyperparameters that we set as shown in Table B.3. The predictions can be derived directly from the representation (in-processing) or, as in our case, by training a classifier on a transformed dataset (pre-processing). By using this method, we expect to find a latent representation of re-employment chances that does not depend on gender.

- **Equalized Odds Postprocessing (EOP) [24].** The EOP technique is a method for achieving fairness by adjusting the predictions of the Machine Learning (ML) model, rather than the input data. The method learns a derived classifier that solves an optimization problem that both maximizes prediction accuracy and satisfies Equalized Odds, which requires the FPR and the True Positive Rate (TPR) to be equal across groups. The predictions are adjusted by setting a different threshold to each group based on the sensitive attribute.

  In the context of resource allocation, aiming for Equalized Odds would imply that differences that exist in the observed data will still be present in the predictions, i.e., if the original data shows higher chances of

re-employment for women, the model would more likely assign a positive label to women. However, we require that the error rates should be the same for both genders, meaning that women who are eligible for support are equally likely to receive it as men, and similarly that men and women who do not need support are as likely not to receive it.

4 RESULTS

4.1 Performance comparison

As shown in Table 1, we find that the observed AUC scores for all (original and adjusted) models are in the range [0.63, 0.65], which is consistent with the results of comparable systems [15, 27]. The AMAS model achieved an accuracy of 0.67, which is only slightly lower than the LR model's accuracy value of 0.69 and thus noteworthy given that the AMAS model was applied to our data without retraining. However, this implies that in practice about one-third of young job seekers would be subject to misclassification.

Table 1. Prediction performance.

Model	Performance metrics				
	AUC	Accuracy	Precision	Recall	F1 Score
AMAS	0.65	0.67	0.50	0.15	0.23
LR	0.64	0.69	0.55	0.31	0.40
RW_{LR}	0.63	0.66	0.48	0.22	0.30
LFR_{LR}	0.63	0.64	0.43	0.29	0.35
EOP_{AMAS}		0.64	0.38	0.13	0.20
EOP_{LR}		0.65	0.45	0.28	0.35

Note. All values (except for AUC) were obtained at the classification threshold $t = 0.66$. AUC scores could not be obtained for EOP as the method did not change model scores.

On our test data, the AMAS model yields 50% precision, which is significantly less than the 73% precision reported in the documentation of the young adults model [22, p. 67]. We obtain similar scores when applying the AMAS model to the complete data set. From the job seekers' perspective, this is not sufficient, as half of the job seekers with a positive prediction are actually not re-employed and would therefore be eligible for support measures. For the LR model, precision is slightly higher with a score of 0.55. When considering the PES objective by interpreting recall scores, the AMAS model only achieves a score of 0.15 on our test dataset, thus allocating resources to 85% of successfully re-employed individuals. Looking at the LR model, about one-third (0.31) of re-employed job seekers are identified as such, hence about two-thirds (0.69) are still misclassified. This result is not desirable under the PES objective of cost-efficient distribution of scarce resources. In direct comparison, the AUC scores for the AMAS model and the LR model are very similar, but all threshold-related performance values are slightly better for the latter.

With respect to the idea that the model could also classify job seekers under the assumption that no support is provided, we recalculated model performance only for those instances in our dataset that did not participate in AMS-funded labor market programs. As shown in Table 1, this analysis leads to slightly better AUC and recall scores for the AMAS model, but decreases accuracy and precision.

Two pre-processing bias mitigation techniques were applied to the dataset. For RW_{LR}, we see slightly lower scores over performance metrics compared to the two models that were not corrected for fairness. The LFR_{LR} model performs

worse than all previously mentioned models on AUC, accuracy and precision, but achieves better recall and F1 score than RW_{LR}. This could be an argument for the PES to consider the LFR over the RW approach and accept (potentially) lower values for accuracy and precision for the purpose of reducing costs. Overall, without considering the impact on fairness, we can conclude that the pre-processing bias mitigation strategies have a negative impact on performance.

When comparing performance of the EOP_{AMAS} model that uses EOP as a post-processing method to improve fairness of the AMAS model, it can be seen that changing the predictions leads to the worst performance compared to all other algorithms under study. For the EOP_{LR} model, performance is only slightly worse after correcting the predictions of the initial LR model. Thus, using the EOP method to correct for bias in our use case leads to a drop in performance, which in turn can cause undesirable effects.

4.2 Fairness evaluation

We next present the results of the fairness evaluation for the models predicting the chances for re-employment with respect to the sensitive attribute `gender`. All fairness metrics were computed with the classification threshold set to $t = 0.66$. For each metric, positive values indicate a preference for the protected group, which in our case is women.

Table 2. Fairness metrics.

Model	Fairness metrics				
	Statistical Parity Diff.	Disparate Impact Scaled	Equal Opportunity Diff. (TPR)	Equal Opportunity Diff. (FPR)	Average Odds Diff.
Observed	0.09	0.23			
AMAS	0.19	0.92	0.19	0.19	0.19
LR	0.30	0.84	0.40	0.23	0.31
RW_{LR}	0.10	0.47	0.10	0.09	0.09
LFR_{LR}	0.09	0.33	0.14	0.06	0.10
EOP_{AMAS}	0.02	0.14	−0.02	0.04	0.01
EOP_{LR}	0.03	0.12	−0.01	0.03	0.01

Table 2 shows that group differences in base rates are present in our test dataset with an observed value of 0.09 for SPD and 0.23 for DIS. Since both values are positive, we can derive that in our test dataset, the share of actual positive outcomes is greater in the protected group, i.e., women, than in the unprotected group, i.e., men. This implies that in our use case, the share of women being re-employed (and thus should not receive support measures) is greater than the share of men. Predicting re-employment chances by means of the AMAS model, the SPD increases to 0.19 and the DIS increases to 0.92. This implies that women are (even) more likely than men to be classified as re-employed when registering with AMS, in turn reducing their odds of receiving support. As our LR model attempts to learn associations between the given attributes, it also reinforces the preference for females, with an SPD of 0.30 and a DIS of 0.84.

For the error-based metrics, we see that the AMAS model is better at classifying women who actually found a job as positive (TPR) than men with a difference of 0.19. This in turn means that among men, more cases were misclassified as not re-employed and thus more men would receive support without needing it. However, the difference in FPR of 0.19 indicates that the rate of women being incorrectly assigned to the group with high re-employment chances ($\hat{Y} = 1$), when in fact they would have needed support, is greater than that of men.

This gender inequality is even more pronounced in the predictions of the LR model with preference for female instances over all fairness metrics. But, similar to the AMAS model, the positive value of EOD considering FPR shows that while more re-employed females are classified as such, the model also tends to misclassify females as positive to a greater extent than males.

Applying RW_{LR} shows that although differences between the groups could not be completely balanced, the procedure improved the fairness metrics. SPD, DIS and error-based metrics achieve values closer to optimal than the AMAS or the LR model, respectively. The LFR_{LR} model even achieves a SPD of 0.09 which is equal to the observed value for this metric, thus retaining the differences that exist in the test dataset. DIS as a ratio measure, however, still indicates a higher ratio of positively labeled females. As for RW_{LR}, LFR_{LR} was able to improve fairness with respect to EOD compared to the non-mitigated model. However, whereas LFR_{LR} leads to a greater difference in TPR favoring female instances, the difference in FPR is smaller compared to the RW_{LR} method.

Table 2 further shows that using EOP to correct the predictions of the AMAS model outperforms not only the non-mitigated model, but also the LR in combination with both pre-processing de-biasing techniques. The SPD has decreased to 0.02, and the DIS to 0.14. Further, the value of -0.02 for EOD_{TPR} indicates a shift in the TPR difference toward males. Generally, correcting for bias using the EOP_{AMAS} and EOP_{LR} methods yield the best results in terms of the selected fairness metrics, as the error-based metrics are close to optimal. Although neither the independence criterion nor the separation criterion can be fully satisfied, EOP was able to considerably reduce the differences between groups for both prediction models.

4.3 Robustness of bias mitigation

As can be seen from the results reported, the decision to set the classification threshold at $t = 0.66$ shows that precision and especially recall take on values that do not seem optimal for practice. The objective of the PES to save costs would not be sufficiently met, and the concerns of job seekers to actually receive measures when needed would not be fulfilled consistently. We can use the F1 score as a measure to reflect these two perspectives and study its dependence on the threshold t, along with the robustness of bias mitigation strategies.

Figure 1 shows that for both the AMAS and LR model, higher F1 scores can be obtained with lower, less restrictive thresholds. Applying EOP to the predictions obtained by the AMAS model leads to a tolerable drop in performance across thresholds. Whereas up to the threshold of 0.3 we do not see clear differences between the original and the mitigated model, this changes for thresholds > 0.3 with the dashed line (mitigated model) being able to decrease biases consistently across metrics and thresholds. This is also true when applying the EOP method to the predictions of the LR model, although here some fluctuation in DIS occurs at higher threshold values.

Overall, we see that in our case study bias mitigation algorithms are able to decrease differences between gender in predicting re-employment chances. Even though the classification threshold can have an impact on the results of the performance and fairness evaluation, no clear opposing trend of the two can be observed. Further, our analysis supports the findings from Friedler et al. [21] that fairness improvements often tend to be visible over multiple metrics. Although the fairness criteria independency and separation contradict each other in theory [13, 30], we show that bias mitigation strategies are able to reduce group differences with respect to both criteria.

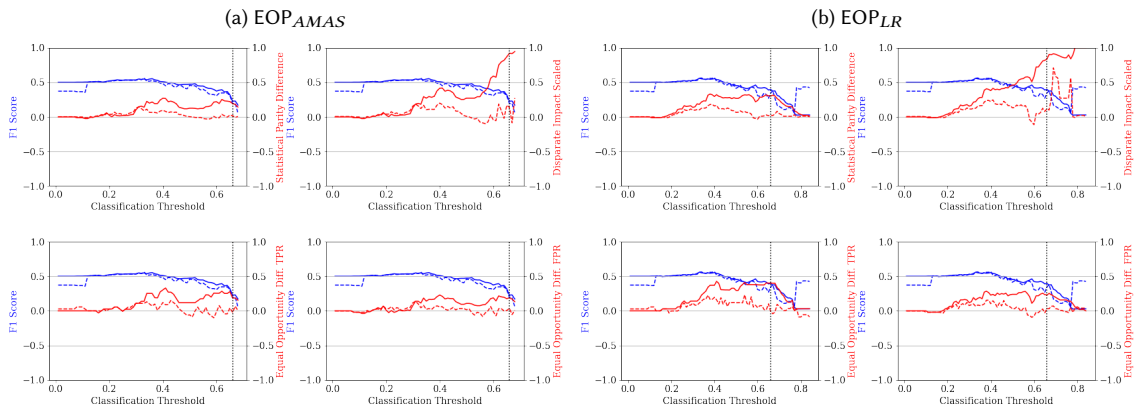

Fig. 1. F1 score and fairness metrics before mitigation (solid line) and after applying EOP (dashed line). The threshold t = 0.66 is indicated by a dotted line

5 DISCUSSION

5.1 Implications of fairness results

To provide a comprehensive interpretation of our results, it is necessary to supplement the statistical fairness results with their implications for practice. In this study, we have used the fairness measures SPD and DIS to assess whether the model meets the independence criterion, which aims to ensure that the proportion of individuals predicted to find employment in the short-term criterion is the same for men and women.

We observed higher re-employment rates among young women which could be explained by several factors such as level of education, but it requires the involvement of domain experts to gain a deeper understanding of the data and the relationships between attributes. Although we do not presume to answer the question of whether the independence criterion ensures fairness, we still see it as necessary to apply bias mitigation strategies in practice to prevent differences between groups from being amplified. Our results show that de-biasing methods such as RW, LFR, and EOP are able to reduce the values of the corresponding fairness metrics, with a tolerable drop in performance. However, these measures do not take into account actual outcomes and are unable to account for existing disparities.

We therefore also evaluated the separation criterion, which states that the sensitive attribute and the outcome should be statistically independent conditioned on the true outcome. We used the fairness metrics EOD and AOD and aimed for an optimal value of 0. For EOD, this would mean that the model has the same TPR or FPR, respectively, for men and women. The former implies that the model is equally good at correctly identifying individuals who actually found a job. If the groups differ greatly in size, it stands to reason that the model would perform better for the majority in order to make fewer errors overall.

As mentioned earlier, a high TPR is beneficial for the PES, as this would mean that not many individuals are incorrectly classified as being eligible for support measures. However, from the job seekers' perspective, higher cost is in a FP prediction, as this would mean that a person who is not re-employed is incorrectly not receiving measures. In general, striving for equality in model performance solves the problem of randomly assigning instances to the positive class to improve fairness. But the assumption that observed data are representative of a *truth* and that future decisions should

be based on the world as it is also has shortcomings. If we assume that, as we saw in our dataset, women tend to have higher re-employment rates than men, using a model that assigns more women to the positive class, which results in them not receiving support, might overlook the fact that the differences could be due to women taking more short-term jobs. In the long term, this could reinforce existing gender differences on the labor market.

We were able to show that all bias mitigation strategies obtained acceptable results with respect to the selected fairness metrics. Although the initial models did not perform equally well for men and women, the differences were reduced by applying de-biasing techniques. As mentioned earlier, overall model performance decreased only slightly and thus does not conflict with the goal of improving fairness in our use case. This allows us to conclude that the methods presented can offer a meaningful contribution to practice.

5.2 Limitations and further research

Our modeling of the AMAS system depends on a reconstructed model that has been disclosed to the public while detailed information on the construction of the original model is lacking. Therefore, an extended fairness audit of the complete system would require disclosure of more details and, ideally, access to the data on which the publications are based. This limited access to data or model descriptions is one of the challenges faced by researchers studying the socio-technical impact of algorithmic systems. To bridge the gap between research and fairness-aware machine learning in practice, we encourage policymakers to provide insights and allow observation of systems to contribute to enhancements in tackling discrimination through algorithmic profiling.

Our study is limited to the investigation of bias based on `gender` as a binary sensitive attribute, which falls short to include people who do not ascribe themselves to binary gender categories. Future research should expand to other sensitive attributes such as citizenship of the job seekers. Additionally, there is a need to investigate the effects of bias mitigation on intersectional discrimination, which considers unequal treatment on multiple grounds simultaneously, as proposed by Morina et al. [43].

Furthermore, all fairness criteria considered in this study have weaknesses with respect to their basic assumptions of the relationships between the protected attribute and the predicted outcome. They are observational criteria, i.e., they only take into account the joint distribution of the features, the protected attribute, the classifier, and the outcome. In our study, the actual relations between the given features, their modification, their differences between groups, and their importance for the prediction were not further studied. Therefore, incorporating causality in the assessment of fairness [29, 36, 39] may be a useful extension to provide further insights for the socio-technical impact assessment.

In addition to the fairness-enhancing interventions presented here, we also applied the Reject Option Classification (ROC) method proposed by [26]. ROC takes into account uncertainty in prediction, but did not achieve any result on our data due to a small number of data points around the classification threshold of 0.66. Therefore, this method was not able to mitigate bias in our use case.

Our findings have demonstrated that almost all of the bias mitigation strategies we employed were able to improve the statistical fairness measures (see subsection 4.2). However, we were not able to directly identify which instances or parameters were modified as a result of the mitigation process. This lack of transparency poses a challenge for interpreting the results and understanding the underlying mechanisms of the mitigation algorithms. To address this limitation, we argue that fairness-enhancing interventions should also provide a way to make the effects of their application visible. This aligns with the growing body of research that aims to combine fairness and explainability in ML (e.g., [23]).

Lastly, the methods presented in this paper provide a technical approach to reducing bias in algorithmic profiling that is necessary, but not sufficient, to overcome the problem of discrimination by data-driven systems. Neither the environment in which the system is used is stable, nor are the characteristics of the individuals used as predictors. If the context in which decisions are made remains unfair, any bias mitigation technique may have limited impact. Therefore, fairness interventions must go beyond the algorithmic system and be considered from a broader perspective. This includes revisiting the processes in the socio-technical environment, such as data collection practices and data selection. With respect to the AMAS case, for example, the selected variables provide limited agency for the job seekers to improve their IC score. Further, integrating domain knowledge into the decision-making process is critical to address fundamental inequities before a fair model is applied. Since the conceptualization of fairness is multidisciplinary, its implementation into technical processes should be discussed by interdisciplinary teams.

6 CONCLUSION

This study provides a socio-technical perspective on how to deal with unequal treatment in a consequential algorithmic profiling setting. As one of the first, we empirically assess the fairness of the AMAS system on a novel real-world dataset of young job seekers from Austria. Next to observing gender-specific error patterns, we were able to show that the performance of the algorithm was not significantly affected by applying bias mitigation strategies. Furthermore, the quantified gender differences could be reduced with respect to both outcome-based and error-based fairness metrics. By replicating an algorithm to be used in practice we provide insights into the importance of fairness audits and how different stakeholder perspectives can conflict. We point out that addressing the issue of fairness is complex and requires more than meeting quantitative fairness benchmarks. The discussion of how society deals with the ethical challenges that are introduced or reinforced by the use of algorithms must involve interdisciplinary perspectives and, most importantly, the people who are affected by these issues. Although we highlight critical aspects of the use of algorithmic profiling of job seekers in this paper, we are optimistic that the well-intentioned use of these techniques, combined with awareness of their consequences, can be a helpful tool and might even have a positive effect on equal opportunities, since discrimination is now visible and can be addressed. Since algorithms require to explicitly formulate objectives, disparities that have implicitly existed for a long time are now subject to debates. To conclude, we hope that our study will contribute to fairness-enhancing interventions not only being studied by academics, but also being used in practice, thus helping to avert possible negative consequences of algorithmic profiling.

REFERENCES

[1] Sray Agarwal and Shashin Mishra. 2021. *Responsible AI: Implementing Ethical and Unbiased Algorithms.* Springer. https://doi.org/10.1007/978-3-030-76860-7

[2] Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. Algorithmic profiling of job seekers in Austria: How austerity politics are made effective. *Frontiers in Big Data* 3 (2020).

[3] Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. Der AMS-Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Retrieved Feb 03, 2023 https://epub.oeaw.ac.at/0xc1aa5576_0x003bfdf3.pdf.

[4] Kin Andersson. 2015. Predictors of re-employment: A question of attitude, behavior, or gender? *Scandinavian Journal of Psychology* 56, 4 (2015), 438–446.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23, 2016 (2016), 139–159.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities.* fairmlbook.org. http://www.fairmlbook.org.

[7] Joachim Baumann, Corinna Hertweck, Michele Loi, and Christoph Heitz. 2022. Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics. Retrieved Nov 29, 2022 from arXiv, https://arxiv.org/abs/2206.02897.

[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder

Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[9] Sebawit G Bishu and Mohamad G Alkadry. 2017. A Systematic Review of the Gender Pay Gap and Factors That Predict It. *Administration & Society* 49, 1 (2017), 65–104.

[10] Ludwig Bothmann, Kristina Peters, and Bernd Bischl. 2022. What Is Fairness? Implications For FairML. Retrieved Jun 2, 2022 from arXiv, https://arxiv.org/abs/2205.09622.

[11] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18. https://doi.org/10.1109/ICDMW.2009.83

[12] Florian Cech, Fabian Fischer, Soheil Human, Paola Lopez, and Ben Wagner. 2019. Dem AMS-Algorithmus fehlt der Beipackzettel. Retrieved Nov 27, 2022 from futurezone, https://futurezone.at/meinung/dem-ams-algorithmus-fehlt-der-beipackzettel/400636022.

[13] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[14] Andreas Czák. 2019. Das Problem mit dem AMS-Algorithmus. Retrieved Nov 27, 2022 from epicenter.works, https://epicenter.works/content/das-problem-mit-dem-ams-algorithmus.

[15] Sam Desiere, Kristine Langenbucher, and Ludo Struyven. 2019. Statistical profiling in public employment services: An international comparison. *OECD Social, Employment and Migration Working Papers, No. 224* (2019). https://doi.org/https://doi.org/10.1787/b5e5f16e-en

[16] Sam Desiere and Ludo Struyven. 2021. Using artificial intelligence to classify jobseekers: the accuracy-equity trade-off. *Journal of Social Policy* 50, 2 (2021), 367–385.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[18] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic Fairness Datasets: The Story so Far. *Data Mining and Knowledge Discovery* 36, 6 (nov 2022), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

[19] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259—-268. https://doi.org/10.1145/2783258.2783311

[20] Pedro Rubim Borges Fortes. 2020. Paths to digital justice: Judicial robots, algorithmic decision-making, and due process. *Asian Journal of Law and Society* 7, 3 (2020), 453–469.

[21] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 329–338. https://doi.org/10.1145/3287560.3287589

[22] Jutta Gamper, Günter Kernbeiß, and Michael Wagner-Pinter. 2020. Das Assistenzsystem AMAS. Zweck, Grundlagen, Anwendung. Retrieved Nov 27, 2021 from AMS Forschungsnetzwerk, https://ams-forschungsnetzwerk.at/pub/13045.

[23] Przemyslaw A Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying fairness and explainability in supervised learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1905–1916.

[24] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. Retrieved Dec 3, 2021 from arXiv, https://arxiv.org/abs/1610.02413.

[25] Jürgen Holl, Günter Kernbeiß, and Michael Wagner-Pinter. 2018. Das AMS-Arbeitsmarktchancen-Modell. Retrieved Jan 16, 2022 from AMS Forschungsnetzwerk, https://ams-forschungsnetzwerk.at/pub/12630.

[26] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929. https://doi.org/10.1109/ICDM.2012.45

[27] Christoph Kern, Ruben L Bach, Hannah Mautner, and Frauke Kreuter. 2021. Fairness in algorithmic profiling: A German case study. Retrieved Nov 8, 2021 from arXiv, https://arxiv.org/abs/2108.04134.

[28] Amir E Khandani, Adlar J Kim, and Andrew W Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34, 11 (2010), 2767–2787.

[29] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems* 30 (2017).

[30] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. Retrieved Feb 9, 2022 from arXiv, https://arxiv.org/abs/1609.05807.

[31] Martin Kocher. 2021. Parlamentarische Anfragebeantwortung: Einsatz des AMS-Algorithmus. Retrieved Feb 03, 2023 https://www.parlament.gv.at/dokument/XXVII/AB/7065/imfname_994537.pdf.

[32] John Körtner and Giuliano Bonoli. 2021. Predictive Algorithms in the Delivery of Public Employment Services. Retrieved Dec 27, 2022 https://osf.io/j7r8y/download.

[33] Joshua A Kroll, Joanna Huey, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G. Robinson, David G Robinson, and Harlan Yu. 2017. Accountable algorithms. *University of Pennsylvania Law Review* 165, 3 (2017), 633–705.

[34] Matthias Kuppler, Christoph Kern, Ruben Bach, and Frauke Kreuter. 2022. From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology* 7 (2022). https://doi.org/10.3389/fsoc.2022.883999

[35] Matthias Kuppler, Christoph Kern, Ruben L Bach, and Frauke Kreuter. 2021. Distributive justice and fairness metrics in automated decision-making: How much overlap is there? Retrieved Nov 8, 2021 from arXiv, https://arxiv.org/abs/2105.01441.

[36] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems* 30 (2017).

[37] Julian Lamont and Christi Favor. 2017. Distributive Justice. In *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[38] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.

[39] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. Retrieved Nov 26, 2022 from arXiv, https://arxiv.org/abs/1805.05859.

[40] Paola Lopez. 2019. Reinforcing intersectional inequality via the AMS algorithm in Austria. In *Critical Issues in Science, Technology and Society Studies. Conference Proceedings of the th STS Conference (Graz: Verlag der Technischen Universität)*. 1–19.

[41] Artan Loxha, Matteo Morgandi, et al. 2014. Profiling the unemployed: a review of OECD experiences and implications for emerging economics. *Social Protection and labor discussion paper* 1424 (2014).

[42] Marco Marabelli, Sue Newell, and Xinru Page. 2018. Algorithmic Decision-Making in the US Healthcare Industry. *Presented at IFIP 8.2, San Francisco, CA* (2018).

[43] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. 2019. Auditing and achieving intersectional fairness in classification problems. Retrieved Nov 11, 2022 from arXiv, https://arxiv.org/abs/1911.01468.

[44] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *Comput. Surveys* 55, 3 (2022), 1–44. https://doi.org/10.1145/3494672

[45] Glenda Quintini and Danielle Venn. 2013. Back to work: Re-employment, earnings and skill use after job displacement. *OECD* (2013).

[46] Nadia Steiber, Monika Mühlböck, and Bernhard Kittel. 2015. Jung und auf der Suche nach Arbeit in Wien: Eine deskriptive Analyse von AMS-Zugängen im Alter von 18 bis 28 Jahren. Retrieved Nov 11, 2021 from the Institutional Repository of the Institute for Advanced Studies, https://irihs.ihs.ac.at/id/eprint/4733. https://doi.org/10.13140/RG.2.1.4650.8248

[47] Nadia Steiber, Monika Mühlböck, Stefan Vogtenhuber, and Bernhard Kittel. 2017. Jung und auf der Suche nach Arbeit in Wien: Beschreibung des JuSAW-Paneldatensatzes und Analysen von Verläufen zwischen den beiden Umfragezeitpunkten. Endbericht Modul 2. Retrieved Nov 11, 2021 from the Institutional Repository of the Institute for Advanced Studies, https://irihs.ihs.ac.at/id/eprint/4734.

[48] András Szigetvari. 2018. AMS bewertet Arbeitslose künftig per Algorithmus. Retrieved Nov 27, 2022 from https://www.derstandard.at/story/2000089095393/ams-bewertet-arbeitslose-kuenftig-per-algorithmus.

[49] Marius Wilk. 2019. Auskunft zum Arbeitsmarktchancen Assistenz-System des AMS. Retrieved Feb 03, 2023 https://epicenter.works/sites/default/files/ams_anfragebeantwortung_vom_16.08.2019_bezgl._ams_algorithmus.pdf.

[50] Martin C Yu and Nathan R Kuncel. 2020. Pushing the limits for judgmental consistency: Comparing random weighting schemes with expert judgments. *Personnel Assessment and Decisions* 6, 2 (2020). https://doi.org/10.25035/pad.2020.02.002

[51] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning* 28, 3 (2013), 325–333.

[52] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology* 32, 4 (2019), 661–683. https://doi.org/10.1007/s13347-017-0293-z

[53] Eva Zschirnt and Didier Ruedin. 2016. Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies* 42, 7 (2016), 1115–1134.

A PRECISION RECALL CURVES

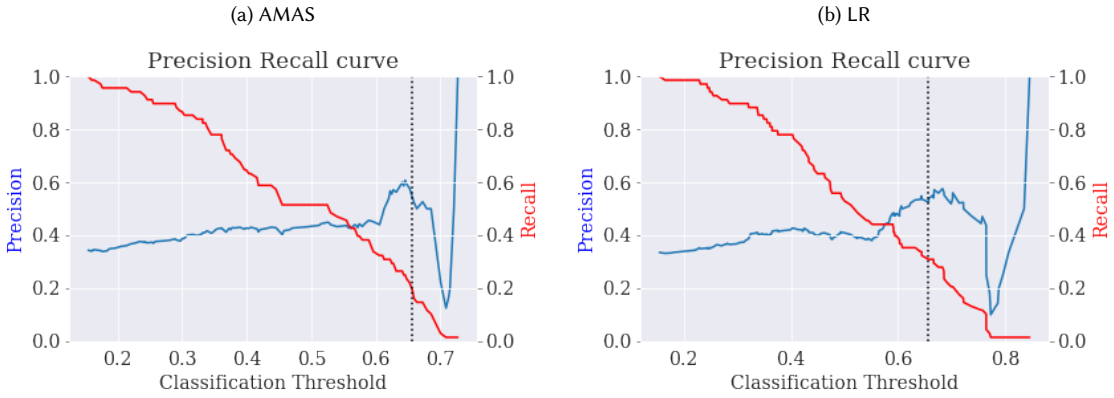

Fig. A.1. Precision and recall curves. The threshold $t = 0.66$ is indicated by a dotted line

B TABLES

Table B.1. Confusion matrix with implications.

		Actual re-employment	
		Positive	Negative
Predicted re-employment	Positive - no support	**True Positive.** A person that is actually being re-employed is predicted as such and thus correctly does not receive support.	**False Positive.** A person that is actually not being re-employed is wrongly predicted as positive and thus wrongly does not receive support.
	Negative - support	**False Negative.** A person that is actually being re-employed is wrongly predicted as negative and thus wrongly receives support.	**True Negative.** A person that is actually not being re-employed is predicted as such and thus correctly receives support.

Table B.2. AMAS coefficients.

Variable	Base value		Coefficient[1]
(Intercept)			−0.30
Gender	Male	Female	0.20
Age	< 20	20-24	0.29
Education	Grade School	Vocational school	1.10
		High school, university	0.93
Obligations of care	No	Yes	−1.05
Regional labor market	Type 1	Type 2	−0.24
		Type 3	−0.40
		Type 4	−0.34
		Type 5	−0.71
Health impairment	No	Yes	−0.82
Occupation	Service	Production	−0.07
Frequency	0 cases	1 case	0.03
		min. 1 case in 2 intervals	−0.04
		min. 1 case in 3 or 4 intervals	−0.13
Measurements claimed	0	min. 1 supportive	−0.48
		min. 1 educational	−0.31
		min. 1 subsidized employment	−0.13
Milestone	0		−0.15

[1]Values are obtained by taking $\log_{10}(OR)$ with OR values reported by Gamper et al. [22, p. 41]

Table B.3. Learning Fair Representations parameters.

Parameter	Value
unprivileged_groups	{'gender_F': 1}
privileged_groups	{'gender_F': 0}
k	5
A_x	0.1
A_y	1.0
A_z	2.0
maxiter	15000
maxfun	15000

Note. k corresponds to the number of prototypes. Further, A_x denotes an input reconstruction quality term weight, A_y an output prediction error and A_z a fairness constraint term weight.

Table B.4. Weights obtained from Reweighing.

Condition	Reweighing values		
	$P(A) * P(Y)$	$P_{obs}(A \wedge Y)$	W
$A = f, Y = 0$	$0.403 * 0.660 = 0.266$	0.241	1.104
$A = f, Y = 1$	$0.403 * 0.340 = 0.137$	0.162	0.846
$A = m, Y = 0$	$0.597 * 0.660 = 0.394$	0.420	0.938
$A = m, Y = 1$	$0.597 * 0.340 = 0.203$	0.177	1.147

Note. Let $P(A)$ be the ratio of instances with the specified value of A, and $P(Y)$ respectively. Note that in our example, $A = f$ represents female instances and $A = m$ represents male instances, respectively. Further, $P_{obs}(A \wedge Y)$ is the observed ratio of instances that fulfill the corresponding feature combination. The weight of each combination is calculated by $W = (P(A) * P(Y))/P_{obs}(A \wedge Y)$.

C BIAS MITIGATION METHODS

C.1 Reweighing [11]

The method of *Reweighing* was proposed by Calders et al. [11]. In this method, each training instance is assigned a weight based on the frequency counts of the protected attribute and the actual outcome. The underlying idea is that if the dataset D was unbiased, i.e., Y is statistically independent of A, then the probability of the joint distribution would be the product of the probabilities as follows:

$$\begin{aligned} P_{expec}(A = a \wedge Y = y) =& P(A = a) \times P(Y = y) \\ =& \frac{|\{A \in D | D(A) = a\}|}{|D|} \times \frac{|\{Y \in D | D(Y) = y\}|}{|D|}, \qquad a, y \in \{0, 1\}, \end{aligned}$$

where $D(A) = a$ are those elements which have the attribute a and $D(Y) = y$, respectively.

In reality, however, datasets often contain biases that result in an observed probability defined as:

$$P_{obs}(A = a \wedge Y = y) = \frac{|\{A, Y \in D | D(A) = a \wedge D(Y) = y\}|}{|D|}, \qquad a, y \in \{0, 1\}$$

To obtain the weights for any combination of the sensitive attribute and outcome, we then compute the fraction of the expected probability and the probability resulting from the observed data, that is:

$$W(X) = \frac{P_{expec}(A = a \wedge Y = y)}{P_{obs}(A = a \wedge Y = y)}, \qquad a, y \in \{0, 1\}$$

By incorporating these weights into the training process, those instances that were disadvantaged (favored) receive higher (lower) weights to compensate for the bias. The corresponding weights for the different groups our dataset are shown in Table B.4.

C.2 Learning Fair Representations [51]

In order to ensure independency between the prediction and the sensitive attribute, Zemel et al. [51] propose *Learning Fair Representations*, a method that creates a latent representation of the data that retains all necessary information about an individual, but obfuscates the group membership derived from a predicted attribute. To formalize this approach, we follow the notation from Zemel et al. [51]. Let X denote a dataset of individuals, where each $x \in X$ is a D-dimensional vector, and X_{train} a training set of individuals. Assume we have access to the protected attribute A, which takes the value u for members of the unprotected group and p for members of the protected group. Let $X^{\mathrm{u}} \subset X$, ${X^{\mathrm{u}}}_{train} \subset X_{train}$ denote the subset of instances (from the whole dataset and the training set, respectively) that are members of the unprotected group, i.e., $A = u$. Accordingly, we denote the subset of instances that are members of the protected group, i.e., $A = p$, as X^{p}, ${X^{\mathrm{p}}}_{train}$. We further introduce Z, a multinomial random variable, where each of the K values represents one of the intermediate set of "prototypes". Given these prototypes, we can then derive a vector v_K for each prototype in the same space as the individuals $x \in X$, where $x = (x_1, ..., x_d)$. We denote d as a distance measure on X and follow the definition by Zemel et al. [51]:

$$d(x_n, v_k, \alpha) = \sum_{i=1}^{D} \alpha_i (x_{n_i} - v_{k_i})^2$$

This distance function allows a different level of impact for each input feature and uses α_i to denote an individual weight parameter for each feature dimension.

LFR aims to learn a mapping that encodes the data as well as possible, but has no information on the sensitive attribute. This constraint follows the notion of Statistical Parity (see subsection 3.3), since it requires that the probability for a random element from X^{u} and a random element from X^{p} map to a given prototype is equal. This can be formulated as:

$$P(Z = k|x^{\mathrm{u}} \in X^{\mathrm{u}}) = P(Z = k|x^{\mathrm{p}} \in X^{\mathrm{p}}), \qquad \forall k \in \{1, ..., K\}$$

As we defined prototypes to be points in the input space, given a set of prototypes we can induce a natural probabilistic mapping from X to Z via the softmax:

$$P(Z = k|x) = \exp(-d(x, v_k))/\sum_{j=1}^{K} \exp(-d(x, v_j))$$

With the three objectives of (1) obfuscating A (L_z), (2) preserving information in X (L_x) and (3) achieving high classification accuracy (L_y), the LFR model aims to minimize the following objective function:

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

For more information on the objective functions, we refer the interested reader to Zemel et al. [51]. Governing the trade-offs, we can set a fairness constraint term weight A_z, an input reconstruction quality term weight A_x and an output prediction error A_y. For these three hyperparameters as well as for the number of prototypes k, we set the values as listed in Table B.3. Predictions can be derived from the representation directly (in that case, LFR would be used in-processing) or, as in our case, by training a classifier on the transformed dataset. By using this method, we expect to find a latent representation of re-employment chances that does not depend on gender.

C.3 Equalized Odds Postprocessing [24]

In their paper on error-based fairness metrics, Hardt et al. [24] present a post-processing method that modifies predictions to satisfy fairness constraints. Their *Equalized Odds Postprocessing* technique learns a derived classifier that in case of a binary predictor gets as input the predicted outcome $\hat{Y}$, the actual outcome Y and the value of the sensitive attribute A. Aiming for Equalized Odds (see subsection 3.3), here denoted as $\gamma_a(\hat{Y})$ where $A = a \in \{0, 1\}$, the method first considers the convex hull, i.e., the set of all convex combinations, of four vertices, defined as:

$$P_a(\hat{Y}) \stackrel{\text{def}}{=} \text{convhull}\{(0, 0), \gamma_a(\hat{Y}), \gamma_a(1 - \hat{Y}), (1, 1)\}$$

The authors further show that the optimal derived predictor $\tilde{Y}$ that yields Equalized Odds can be formulated by the following optimization problem [24]:

$$\begin{aligned} \min_{\tilde{Y}} \quad & \mathbb{E}\ell(\tilde{Y}, Y) \\ s.t. \quad & \forall a \in 0, 1 : \gamma_a(\tilde{Y}) \in P_a(\hat{Y}) \\ & \gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y}) \end{aligned}$$

In the case of a binary classification problem, the above optimization problem is linear. For the extension of this idea to deriving a non-discriminating predictor from a score function, we refer the interested reader to Hardt et al. [24].

By allowing different thresholds for each group of the protected attribute, EOP solves an optimization problem that both maximizes prediction accuracy and satisfies Equalized Odds. In the context of resource allocation, aiming for Equalized Odds would imply that differences that exist in the observed data will still be present in the predictions, i.e., if the original data shows higher chances of re-employment for women, the model would more likely assign a positive label to women. This implies that the error rate should be the same for both genders, meaning that women who are eligible for support are equally likely to receive it as men, and similarly that men and women who do not need support are as likely not to receive it.