Minds and Machines (2025) 35:9
https://doi.org/10.1007/s11023-024-09706-9

Fairness in Algorithmic Profiling: The AMAS Case

Eva Achterhold[1] · **Monika Mühlböck**[2] · **Nadia Steiber**[2] · **Christoph Kern**[1,3]

Received: 31 May 2023 / Accepted: 5 November 2024

Abstract

We study a controversial application of algorithmic profiling in the public sector, the Austrian AMAS system. AMAS was supposed to help caseworkers 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 may form an important building block in revising profiling schemes in the present context.

Keywords Algorithmic profiling · Statistical discrimination · Public employment services · Artificial intelligence · Bias mitigation

✉ Eva Achterhold
mail@acheva.de

[1] LMU Munich, Munich, Germany

[2] University of Vienna, Vienna, Austria

[3] Munich Center for Machine Learning (MCML), Munich, Germany

Published online: 29 January 2025

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 criminal justice (Fortes, 2020), education (Kizilcec & Lee, 2022), public health (Barda et al., 2020; Marabelli et al., 2018) and credit scoring (Khandani et al., 2010). Algorithmic profiling may also be used to allocate severe punitive actions, as demonstrated by controversial applications in the Netherlands which included the prediction-based identification of welfare fraud (van Bekkum & Borgesius, 2021). Algorithmic profiling systems enable the integration of vast amounts of data and thus, on their onset, promise to be more reliable (Barocas et al., 2019), efficient (Lepri et al., 2018), transparent (Zerilli et al., 2019), and accountable (Kroll et al., 2017) 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 (Yu & Kuncel, 2020).

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 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 (Angwin et al., 2016). 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 caseworkers with the decision of allocating service resources to job seekers, in a pilot phase (Holl et al., 2018). The idea was to supplement the caseworker's subjective assessment with a standardized, data-driven evaluation of a person's chances of re-employment. The expected benefits of this process 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 efficiency of the process, such that caseworkers can 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 (Allhutter et al., 2020a).

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 different groups of job seekers. The criterion of 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 (Allhutter et al., 2020a; Cech et al., 2019; Lopez, 2019), journalists (Szigetvari, 2018) and privacy groups (Czák, 2019). 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 (Holl et al., 2018). 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 (Holl et al., 2018). 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 (Bishu and Alkadry, 2017) and ethnicity (Zschirnt and Ruedin, 2016). In 2020, the Austrian Data Protection Authority (DPA) prohibited AMAS, arguing that a legal basis for conducting "profiling" was missing (Kocher, 2021). 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. (2020a) and Linecker (2022) study the AMAS by means of a document analysis and qualitative interviews with PES caseworkers, respectively, an empirical investigation with a focus on fairness metrics and bias mitigation techniques is lacking. Utilizing 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 debiasing 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 (Fabris et al., 2022; Pessach and Shmueli, 2022), we set out to compare debiasing 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 while inducing only a modest drop in performance. Although the choice of the classification threshold affects the performance and fairness metrics, we find that debiasing 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. An algorithmic profiling setting that is considered or used in many countries is prediction-based profiling to identify individuals who are at high risk of Long-Term Unemployment (LTU) (Loxha et al., 2014). 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 Desiere et al. (2019) and Loxha et al. (2014).

In the academic literature on public administration, a lively debate about "digital era governance" (Dunleavy & Margetts, 2023; Dunleavy et al., 2005; Tan & Crompvoets, 2022) and its potential benefits and hazards has evolved. Coining the term "New Public Analytics" (in reference to "New Public Management" describing efficiency-driven reforms of the public sector from the 1980 s onwards), Yeung

(2023) stresses pathologies and dangers of using algorithms in the administration of welfare state politics. Even general proponents of digital governance caution against the use of algorithmic shortcut solutions and automated decision-making on individual cases (Dunleavy and Margetts, 2023). A main strand of research discusses issues of fairness and potential discrimination in algorithmic decisions (Criado & Such, 2019; Yeung, 2019). Broader debates on the distribution of benefits and burdens have long been a topic in philosophy and are closely linked to questions of equality, equity, and justice. The discourse on distributive justice, i.e., the consideration of just allocation of resources among members of a society (Lamont & Favor, 2017), 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) (Baumann et al., 2022; Kuppler et al., 2021). 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. (2022) 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 interventions in the sociotechnical 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 caseworkers, either by relying solely on their expertise or by following rules such as passing a threshold for time in unemployment (Loxha et al., 2014). 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) (Barocas et al., 2019) can be used to evaluate to what extent social groups are treated differently in the decision-making process.

In their study on algorithmic profiling, Kern et al. (2024) applied different classification methods to predict the risk of becoming LTU with German data. They could show that although the statistical models had a similar level of accuracy they had 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). Körtner and Bach (2023) focus on the allocation process directly and demonstrate how the algorithmic allocation of support programs to jobseekers can acknowledge inequalities in baseline (employment) risks while optimizing for the most effective allocation of programs. Zezulka and Genin (2024) show that algorithms that adhere to fairness criteria such as statistical parity and equality of opportunity are not a sufficient criterion to close the gender gap in (un)employment.

Under the criterion of independence, we study whether the chance of receiving support resources depends on the gender of a person. Specifically, we calculate Statistical Parity Differences and an adapted version of Disparate Impact (see Sect. 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 (Andersson, 2015; Quintini & Venn, 2013). An algorithm for guiding the distribution of support measures that meets the independency criterion would not sufficiently account for these differences.

Fair predictions. In algorithmic profiling to allocate public resources, the decisions are informed by predictions 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 metrics (e.g, Equal Opportunity, Equalized Odds) (Barocas et al., 2019).

The separation criterion requires that the error rates of the classifier are equal across groups, in reference to the distribution of the observed outcome. In the context of the allocation 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 (Körtner & Bonoli, 2021). Desiere and Struyven (2021) 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 the 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. In the German context, Bach et al. (2023) and Kern et al. (2024) highlight the importance of modeling choices and show how the set of job seekers that are predicted as being at high risk of LTU differs under different algorithmic profiling schemes.

With respect to the AMAS use case, Allhutter et al. (2020a) provide a systematic document review emphasizing the sociotechnical implications and consequences of the use of the proposed system. Lopez (2019) extends the analysis of the AMAS system by also raising issues of intersectional discrimination, legislation, and the efficiency of the system. Linecker (2022) explores the practical application of the system by PES caseworkers during the trial period. All three works, however, do not provide a data-based fairness assessment of the AMAS system.

The aforementioned studies examining the fairness of algorithmic profiling systems for the unemployed (Bach et al., 2023; Desiere & Struyven, 2021; Kern et al., 2024; Zezulka & Genin, 2024) present the few empirical research efforts in this area. 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. Our work contributes to the literature on fairness of unemployment profiling by applying fairness metrics and bias mitigation strategies to a real-world use case: the AMAS system. Before turning to the methods used in this study, we will briefly discuss 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 into 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 individual and the information about the actual outcome that we know from evaluation data, we can evaluate which individuals have been misclassified, i.e., assigned a predicted outcome that differs from the actual outcome. Misclassification can occur in two ways: First, the algorithm wrongly predicts a negative outcome, which is referred to as a False Negative (FN), and second, the algorithm incorrectly predicts a positive outcome, 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 3 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 this scenario implies considerable social costs, we aim for a low FPR in our performance analysis in order to account for the job seekers' perspective. As another relevant measure that takes FP into account, we use precision (see Sect. 3.3.1), which reflects 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, besides taking into account the FPR, we 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 $t = 0.66$ (Holl et al., 2018). 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 Fig. 2. 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] (Steiber et al., 2015, 2017). 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 (Gamper et al., 2020; Holl et al., 2018). For each individual in our dataset, we created a binary variable which represents the `re-employment` outcome and follows 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 (Allhutter et al., 2020b; Wilk, 2019). 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 model evaluation.

It should be noted that in our data individuals have been subject to case worker-based profiling as the AMAS model was not yet implemented. Thus, individuals may have received support measures between the first and the second interview as assigned by case workers, which in turn can affect model evaluations (Coston et al., 2020). While we are not able to directly control for the influence of support measures between the two interviews with the available data, it does include information about participation in support measures within the last four years prior to the first JuSAW interview. Specifically, about 28% ($n = 187$) of the job seekers in our dataset entered an AMS-funded support measure within this time range. We use this data from the four years prior to the first interview to acknowledge the potential effect of interventions on model evaluation in additional analyses (Table 5 as discussed in Sect. 4.1). Also note that as in the original AMAS model, we use the variable `measures claimed` to account for the impact of prior interventions when predicting re-employment chances (see Table 4).

We closely resemble the AMAS process in our encoding of the predictor variables (Table 4 again). This includes the following pre-processing decisions: Similar to the AMAS, we include `health impairment` as a feature in our analysis. Since we had no information on the health status of the job seeker from the register data, we drew on a variable from the survey data. The participants were asked if they are impaired by any health issue in their daily life with the response options "yes, strongly", "yes", "no" and "prefer not to say". In accordance with the AMAS, we grouped together both "yes" options and set the "prefer not to say" instances to NA (not available). 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 survey information on the number of children of a person. To align our model with the AMAS algorithm, we set the value for `obligations of care` to 0 for all men and to 1 if a person is female and has at least one child.

3.2 Prediction

As mentioned earlier, the procedure for determining an individual's IC score with AMAS is based on a stratification analysis of data on 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 (Gamper et al., 2020; Holl et al., 2018).

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 4).
- **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** $Y \in \{0, 1\}$. Individuals 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** $R \in [0, 1]$, 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.** For a simplified representation of the originally used stratification procedure, coefficients of a logistic regression were published by the AMS (Holl et al., 2018). Based on this, we reconstruct the AMAS model for young adults with the coefficients shown in Table 4.
- **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* (Bellamy et al., 2018).

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. The 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 Sect. 2. All 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 (Andersson, 2015; Quintini & Venn, 2013), 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)** (Dwork et al., 2012). 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)** (Feldman et al., 2015). 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)** (Hardt et al., 2016). 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)** (Agarwal & Mishra, 2021). A classifier's fairness with regard to Equalized Odds can be measured by the AOD, which is defined as follows:

$$\begin{aligned} AOD = \frac{1}{2} \times \big(& (P(\hat{Y} = 1|A = p, Y = 0) - P(\hat{Y} = 1|A = u, Y = 0)) \\ & + (P(\hat{Y} = 1|A = p, Y = 1) - P(\hat{Y} = 1|A = u, Y = 1)) \big) \end{aligned}$$

3.4 Bias Mitigation

In this study, we restrict our selection of bias mitigation techniques to methods that are applicable to categorical inputs and a binary sensitive attribute. Given our focus on scenarios where a classifier, such as the one described in the AMAS example, has been identified as discriminatory after it has been introduced, our analysis focuses on pre- and post-processing techniques (see also Appendix C). However, to ensure a comprehensive analysis of various bias mitigation strategies, we also incorporate fair logistic regression as an in-processing method. We evaluate both the original and mitigated results along the metrics presented in the previous section.

- **Reweighing (RW)** (Calders et al., 2009). 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.

  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 7), we can only apply this method to the LR model but not to the AMAS model.
- **Learning Fair Representations (LFR)** (Zemel et al., 2013). The LFR method aims to learn a new representation of data that is both predictive and fair by removing bias from the input data. 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 a 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, the reconstruction term, and the output prediction error. These weights, as well as the number of prototypes, are hyperparameters that we set as shown in Table 6. 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.

- **Fair Logistic Regression (FLR)** (Zafar et al., 2017). As an in-processing method, we implemented a fair linear logistic regression classifier that aims to balance accuracy with fairness constraints, adhering to the p%-rule. This rule mandates that the positive prediction rate for the unprotected group must be at least 80% of that for the protected, addressing DI. The method adjusts the decision boundary to minimize the covariance between the sensitive attribute and the distance to the decision boundary, thereby reducing potential bias in decision-making.

  This approach maintains the simplicity of the training process, as it does not add complexity to the logistic regression model. By applying this method, we aim to minimize the log-likelihood loss while ensuring the fairness constraints regarding the covariance are satisfied, although this may involve trade-offs in terms of predictive accuracy under certain conditions.
- **Equalized Odds Postprocessing (EOP)** (Hardt et al., 2016). EOP 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 FPR and the TPR to be equal across groups. The predictions are adjusted by setting a different threshold to each group based on the sensitive attribute.

  In this context, Equalized Odds requires 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 to not 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 (Bach et al., 2023; Desiere et al., 2019; Kern et al., 2021). The AMAS model achieved an accuracy of 0.67, which is only slightly lower than the

Table 1 Prediction performance

Model	Performance metrics					
	AUC	Accuracy	Precision	Recall	F1 score	FPR
AMAS	0.65	0.67	0.50	0.15	0.23	0.07
LR	0.64	0.69	0.55	0.31	0.40	0.13
RW_{LR}	0.63	0.66	0.48	0.22	0.30	0.11
LFR_{LR}	0.63	0.64	0.43	0.29	0.35	0.19
FLR	0.56	0.66	0.49	0.24	0.32	0.13
EOP_{AMAS}		0.64	0.38	0.13	0.20	0.11
EOP_{LR}		0.65	0.45	0.28	0.35	0.17

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

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.

On our test data, the AMAS model yields 50% precision, which is significantly lower than the 73% precision reported in the documentation of the AMAS model (Gamper et al. 2020, p. 67). We obtain similar scores when applying the AMAS model on the entire data. However, from the job seekers' perspective, this is not sufficient, as half of those predicted positively are actually not re-employed and would thus be eligible for support measures. Looking at the FPR, among all job seekers that actually do not find a job, the AMAS model mistakenly classifies 7% of them as re-employed. For the LR model, precision is slightly higher, with a score of 0.55 whereas FPR is higher as well (0.13). 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.

In an additional analysis, we recalculated model performance scores only for those individuals in our data that did not participate in AMS-funded labor market programs within the last four years before registering with the AMS. As shown in Table 5, this analysis leads to slightly better AUC and recall scores for the AMAS model, but decreases accuracy and precision. These (modest) performance differences may be caused by noise due to participation in AMS-funded programs, as they affect the re-integration chances that are predicted by the model.

Two pre-processing bias mitigation techniques were applied to the data. For RW_{LR}, we see slightly lower performance scores than the models without fairness adjustments. While RW_{LR} achieves higher values for recall and F1 score compared to the AMAS model, it is also associated with a higher FPR, which is not desirable from the job seekers' perspective. The LFR_{LR} model performs worse than all previously mentioned models on AUC, accuracy, precision and FPR, but achieves better recall and F1 scores 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.

We further trained a fair LR classifier FLR as proposed by Zafar et al. (2017) on our data in order to see whether AMAS-like systems could be explicitly trained for fairness. Our analysis reveals that when accounting for fairness constraints, this classifier achieves 66% accuracy and achieves lower values over all performance metrics compared to the unconstrained LR.

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. 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 shows that group differences in base rates are present in our test data with an observed value of 0.09 for SPD and 0.23 for DIS. Since both values are positive,

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
FLR	0.12	0.53	0.07	0.00	0.03
EOP_{AMAS}	0.02	0.14	−0.02	0.04	0.01
EOP_{LR}	0.03	0.12	−0.01	0.03	0.01

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 who 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. 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 over all fairness metrics. 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. Similar to RW_{LR}, LFR_{LR} is able to improve fairness with respect to EOD compared to the non-mitigated model. However, while LFR_{LR} leads to a greater difference in TPR favoring women, the difference in FPR is smaller compared to the RW_{LR} method.

We employed fair logistic regression FLR as an example of in-processing bias mitigation, optimizing the classifier for accuracy while aiming for perfect fairness. The outcomes with regard to fairness indicate a reduction in SPD and DIS to 0.12 and 0.53, respectively, with the classifier achieving values closer to zero across all other metrics compared to those achieved by pre-processing methods. With only a slight drop of 0.03 in accuracy compared to the LR, the FLR outperforms the unconstrained model with regard to fairness. These results confirm the viability of in-processing methods within the context of this work.

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 debiasing techniques. Although the EOP method mainly aims to correct for error-based metrics, the SPD has decreased to 0.02, and the DIS to 0.14. This illustrates the intercorrelation of fairness metrics as both types of metrics, outcome-based and error-based, take into account the rate of positive outcomes per group. An improvement in one metric can thus lead to an improvement in the other. Furthermore, the negative value 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 mentioned earlier, the decision to set the classification threshold to $t = 0.66$ leads to precision and recall values that do not seem optimal in 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. While we do not see differences between the original and the mitigated model up to a threshold of 0.3, 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 observe that in our case study, bias mitigation techniques are able to decrease gender disparities when 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. Furthermore, our analysis supports the findings from Friedler et al. (2019) that fairness improvements often tend to be visible over multiple metrics. Although the fairness criteria independency and separation contradict each other

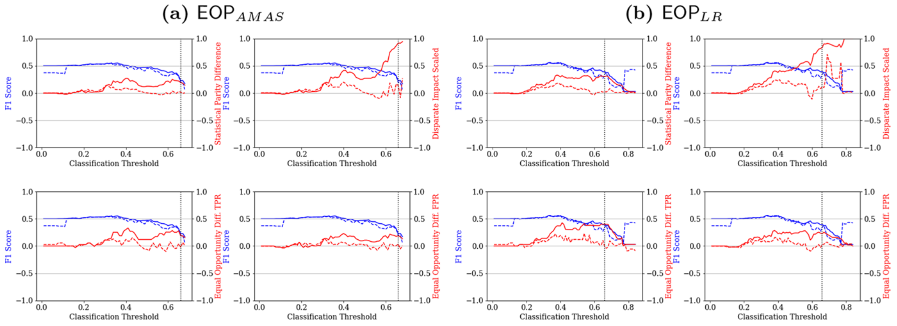

Fig. 1 F1 score and fairness metrics before mitigation (solid line) and after applying EOP (dashed line). The threshold t = 0.66 is indicated by a dotted line

in theory (Chouldechova, 2017; Kleinberg et al., 2016), we show that bias mitigation strategies are able to reduce group differences with respect to both criteria.

5 Discussion

5.1 Implications of Fairness Results

To provide a comprehensive interpretation of our results, it is necessary to connect the statistical fairness results with their implications for practice. In this study, we 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 education, but it requires domain experts to gain a deeper understanding of the underlying mechanisms and relationships between attributes. Although we do not presume to answer the question of whether the independence criterion ensures fairness, we would argue that at least differences between groups should not be amplified by prediction algorithms. Our results show that debiasing 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, measures such as SPD and DIS 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 debiasing techniques. As mentioned earlier, overall model performance decreased only slightly and thus did 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 in practice.

In summary, our results highlight that implementing algorithmic profiling systems in the public sector raises critical questions and can potentially reinforce biases and historical discrimination rather than improving objectivity in decision-making. Similar to Desiere and Struyven (2021) and Kern et al. (2024), we observed disparate error rates across groups of job seekers defined by sensitive personal attributes. Taken together, the findings imply that prediction models in labor market contexts tend to over-amplify differences between groups that may be observable in training data, which in turn can exacerbate discrimination depending on how the predictions are used in the eventual decision-making. Whether error differences in this setting can be robustly mitigated with debiasing techniques is, however, an open question. Although we observe positive results, studies have shown that debiasing can also fail to provide any improvement and that results are highly context-dependent (Agrawal et al., 2021).

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 models are based. This limited access to data or model descriptions is one of the challenges faced by researchers studying the sociotechnical 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 access to systems such that discrimination in algorithmic profiling can be studied and mitigated effectively.

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 consider additional 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. (2019).

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 (Kilbertus et al., 2017; Kusner et al., 2017; Loftus et al., 2018) may be a useful extension to provide further insights for the sociotechnical impact assessment.

The fairness criteria considered in this study all aim at ensuring fairness for members of social groups that are based on protected attributes. When measuring fairness, a valid consideration is who is the subject of this evaluation. Here, we can distinguish between group and individual fairness. Group fairness approaches can typically be applied without detailed external knowledge on the domain of application and without further assumptions (Chouldechova & Roth, 2020). To extend beyond group fairness, Dwork et al. (2012) suggest to focus on individual fairness, stating that similar individuals should be treated similarly. However, assessing similarity is task-specific and need to be determined for each use case separately. This notion of fairness thus requires finding a metric for similarity, which itself is a non-trivial fairness problem (Chouldechova & Roth, 2020; Pessach & Shmueli, 2022). A similar approach to define individual fairness is the idea to not favor less qualified individuals over more qualified individuals (Joseph et al., 2016). Since assessing quality requires knowledge on the true underlying label that is unknown to the algorithm, this definition can be hardly put to practice (Chouldechova & Roth, 2020). Given the limited body of literature on measuring and enhancing individual fairness in practice, we opted to solely focus on group fairness measures in this work but encourage future work on individual fairness implications of algorithmic profiling on the labor market.

As in similar use cases where algorithms are used for risk assessment, the AMAS model is based on observations that have been influenced by the historical decision-making policy. That is, the coefficients we use to reconstruct the AMAS algorithm are based on training data in which participation in active labor market programs can impact an individual's chances of re-employment. Thus, the predicted integration chance does not imply how likely it is for an individual to find employment with respect to the short-term criterion without receiving support measures. Instead, it is conditioned on the continued allocation of support measures by caseworkers. Coston et al. (2020) use counterfactual risk assessment to account for the effects of the intervention in fairness assessment. However, their method requires an outcome that is observable for the control group without intervention. Applying this approach to the AMAS model would not be feasible as it would require withholding measures from individuals solely for the purpose of data collection, which would not be consistent with ethical principles. If only those individuals who would not have received measures based on the caseworkers' decision policy were included in the assessment of re-employment chance, this would introduce a selection bias. Furthermore, the approach requires training a separate model for the counterfactual case, which in turn requires access to the model and the original training data, which is not available in our case study.

In addition to the fairness-enhancing interventions presented here, we also applied the Reject Option Classification (ROC) method proposed by Kamiran et al. (2012). 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 the bias mitigation strategies we employed were able to improve the statistical fairness measures (see Sect. 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., Grabowicz et al. 2022).

Lastly, the methods presented in this paper provide a technical approach to reducing bias in algorithmic profiling which 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 sociotechnical 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 sociotechnical 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.

Appendix A: Precision Recall Curves

See Fig. 2.

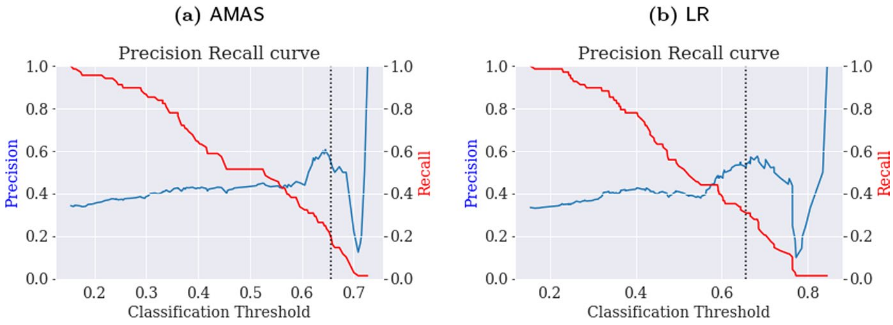

Fig. 2 Precision and recall curves. The threshold $t = 0.66$ is indicated by a dotted line

Appendix B: Tables

See Tables 3, 4, 5, 6, and 7.

Table 3 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 4 AMAS coefficients

Variable	Base value		Coefficient[a]
(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

[a] Values are obtained by taking $\log_{10}(OR)$ with OR values reported by (Gamper et al. 2020, p. 41)

Table 5 Prediction performance of AMAS model conditioned on measures claimed

Data	Performance metrics				
	AUC	Accuracy	Precision	Recall	F1 score
$X_{m=0,1}$	0.65	0.67	0.50	0.15	0.23
$X_{m=0}$	0.68	0.62	0.48	0.22	0.3

We set $m = 1$ to indicate if a person did receive support measures in the past four years or not ($m = 0$). Let X_m describe those instances in our dataset X that fulfill the criterion m. All values (except for AUC) were obtained at the classification threshold $t = 0.66$

Table 6 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

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 7 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

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 $W = (P(A) * P(Y))/P_{obs}(A \wedge Y)$

Appendix C: Bias Mitigation Methods

C.1 Reweighing (Calders et al., 2009)

The method of *Reweighing* was proposed by Calders et al. (2009). 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 7.

C.2 Learning Fair Representations (Zemel et al., 2013)

In order to ensure independency between the prediction and the sensitive attribute, Zemel et al. (2013) 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. (2013). 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 $X^u \subset X$, $X^u_{train} \subset X_{train}$ 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^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. (2013):

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

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 Sect. 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:

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

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. (2013). 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 6. 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 (Hardt et al., 2016)

In their paper on error-based fairness metrics, Hardt et al. (2016) 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 Sect. 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:

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

The authors further show that the optimal derived predictor $\tilde{Y}$ that yields Equalized Odds can be formulated by the following optimization problem (Hardt et al., 2016):

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

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. (2016).

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.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose

References

Agrawal, A., Pfisterer, F., Bischl, B., Buet-Golfouse, F., Sood, S., Chen, J., Shah, S., & Vollmer, S. (2021). Debiasing classifiers: Is reality at variance with expectation? arXiv:2011.02407

Agarwal, S., & Mishra, S. (2021). *Responsible AI: Implementing ethical and unbiased algorithms*. Springer

Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020a). Algorithmic profiling of job seekers in Austria: How austerity politics are made effective. *Frontiers in Big Data*, 3.

Allhutter, D., Mager, A., Cech, F., Fischer, F., & Grill, G. (2020b). Der AMS-Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Retrieved February 3, 2023, from https://epub.oeaw.ac.at/0xc1aa5576_0x003bfdf3.pdf.

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

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, 23*(2016), 139–159.

Bach, R. L., Kern, C., Mautner, H., & Kreuter, F. (2023). The impact of modeling decisions in statistical profiling. *Data & Policy, 5*, e32.

Barda, N., Riesel, D., Akriv, A., Levy, J., Finkel, U., Yona, G., Greenfeld, D., Sheiba, S., Somer, J., Bachmat, E., & Rothblum, G. N. (2020). Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nature Communications, 11*(1), 4439.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. http://www.fairmlbook.org

Baumann, J., Hertweck, C., Loi, M., & Heitz, C. (2022). Distributive justice as the foundational premise of fair ML: Unification, extension, and interpretation of group fairness metrics. Retrieved November 29, 2022 from https://arxiv.org/abs/2206.02897

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *Understanding, and Mitigating Unwanted Algorithmic Bias*, 2

Bishu, S. G., & Alkadry, M. G. (2017). A systematic review of the gender pay gap and factors that predict it. *Administration & Society, 49*(1), 65–104.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops* (pp. 13–18). IEEE

Cech, F., Fischer, F., Human, S., Lopez, P., & Wagner, B. (2019). Dem AMS-Algorithmus fehlt der Beipackzettel. Retrieved November 27, 2022, from https://futurezone.at/meinung/dem-ams-algorithmus-fehlt-der-beipackzettel/400636022

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 5*(2), 153–163.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM, 63*(5), 82–89.

Coston, A., Mishler, A., Kennedy, E. H., & Chouldechova, A. (2020). Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 582–593).

Criado, N., & Such, J. M. (2019). Digital discrimination. *Algorithmic Regulation*, 82–97.

Czák, A. (2019). Das Problem mit dem AMS-Algorithmus. Retrieved November 27, 2022, from https://epicenter.works/content/das-problem-mit-dem-ams-algorithmus.

Desiere, S., Langenbucher, K., & Struyven, L. (2019). Statistical profiling in public employment services: An international comparison. *OECD Social, Employment and Migration Working Papers, No. 224.*

Desiere, S., & Struyven, L. (2021). Using artificial intelligence to classify jobseekers: The accuracy-equity trade-off. *Journal of Social Policy, 50*(2), 367–385.

Dunleavy, P., & Margetts, H. (2023). Data science, artificial intelligence and the third wave of digital era governance. *Public Policy and Administration*. https://doi.org/10.1177/09520767231198737

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2005). New public management is dead-long live digital-era governance. *Journal of Public Administration Research and Theory, 16*(3), 467–494.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226)

Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery, 36*(6), 2074–2152.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259—268).

Fortes, P. R. B. (2020). Paths to digital justice: Judicial robots, algorithmic decision-making, and due process. *Asian Journal of Law and Society, 7*(3), 453–469.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 329–338).

Gamper, J., Kernbeiß, G., & Wagner-Pinter, M. (2020). Das Assistenzsystem AMAS. Zweck, Grundlagen, Anwendung. Retrieved November 27, 2021 from https://ams-forschungsnetzwerk.at/pub/13045

Grabowicz, P. A., Perello, N., & Mishra, A. (2022). Marrying fairness and explainability in supervised learning. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 1905–1916)

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. https://arxiv.org/abs/1610.02413

Holl, J., Kernbeiß, G., & Wagner-Pinter, M. (2018). Das AMS-Arbeitsmarktchancen-Modell. Retrieved January 16, 2022 from https://ams-forschungsnetzwerk.at/pub/12630

Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems*, 29

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining* (pp. 924–929). IEEE

Kern, C., Bach, R., Mautner, H., & Kreuter, F. (2021). Fairness in algorithmic profiling: A German case study. https://arxiv.org/abs/2108.04134

Kern, C., Bach, R., Mautner, H., & Kreuter, F. (2024). When small decisions have big impact: Fairness implications of algorithmic profiling schemes. *ACM Journal on Responsible Computing*

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767–2787.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 30

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The ethics of artificial intelligence in education* (pp. 174–202). Routledge

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. https://arxiv.org/abs/1609.05807

Kocher, M. (2021). Parlamentarische Anfragebeantwortung: Einsatz des AMS-Algorithmus. Retrieved February 03, 2023 https://www.parlament.gv.at/dokument/XXVII/AB/7065/imfname_994537.pdf

Körtner, J., & Bach, R. (2023). Inequality-averse outcome-based matching

Körtner, J., & Bonoli, G. (2021). Predictive algorithms in the delivery of public employment services. Retrieved December 27, 2022 https://osf.io/j7r8y/download

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633–705.

Kuppler, M., Kern, C., Bach, R., & Kreuter, F. (2021). Distributive justice and fairness metrics in automated decision-making: How much overlap is there? https://arxiv.org/abs/2105.01441

Kuppler, M., Kern, C., Bach, R., & Kreuter, F. (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

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30

Lamont, J., & Favor, C. (2017). Distributive justice. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Winter 2017 edition.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611–627.

Linecker, L. E. (2022). Diskriminiert das ams-arbeitsmarktchancen-assistenzsystem?: Mögliche auswirkungen des "ams-algorithmus" auf arbeitssuchende frauen. https://utheses.univie.ac.at/detail/65161

Loftus, J. R., Russell, C., Kusner, M. J., & Silva, R. (2018). Causal reasoning for algorithmic fairness. https://arxiv.org/abs/1805.05859.

Lopez, P. (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)* (pp. 1–19).

Loxha, A., Morgandi, M. (2014). Profiling the unemployed: A review of OECD experiences and implications for emerging economics. *Social Protection and Labor Discussion Paper* (p. 1424).

Marabelli, M., Newell, S., & Page, X. (2018). Algorithmic decision-making in the US Healthcare Industry. *Presented at IFIP 8.2, San Francisco, CA*

Morina, G., Oliinyk, V., Waton, J., Marusic, I., & Georgatzis, K. (2019). Auditing and achieving intersectional fairness in classification problems. Retrieved November 11, 2022 from https://arxiv.org/abs/1911.01468

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys, 55*(3), 1–44.

Quintini, G., & Venn, D. (2013). Back to work: Re-employment, earnings and skill use after job displacement. *OECD*

Steiber, N., Mühlböck, M., & Kittel, B. (2015). Jung und auf der Suche nach Arbeit in Wien: Eine deskriptive Analyse von AMS-Zugängen im Alter von 18 bis 28 Jahren. Institutional Repository of the Institute for Advanced Studies. Retrieved November 11, 2021 from https://irihs.ihs.ac.at/id/eprint/4733.

Steiber, N., Mühlböck, M., Vogtenhuber, S., & Kittel, B. (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. Institutional Repository of the Institute for Advanced Studies. Retrieved November 11, 2021 from https://irihs.ihs.ac.at/id/eprint/4734.

Szigetvari, A. (2018). AMS bewertet Arbeitslose kúnftig per Algorithmus. Retrieved November 27, 2022 from https://www.derstandard.at/story/2000089095393/ams-bewertet-arbeitslose-kuenftig-per-algorithmus

Tan, E., & Crompvoets, J. (2022). *Chapter 1: A new era of digital governance, chapter 1* (pp. 13–49).

van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security, 23*(4), 323–340.

Wilk, M. (2019). Auskunft zum Arbeitsmarktchancen Assistenz-System des AMS. Retrieved February 03, 2023 https://epicenter.works/sites/default/files/ams_anfragebeantwortung_vom_16.08.2019_bezgl._ams_algorithmus.pdf

Yeung, K. (2019). Why worry about decision-making by machine? In: *Algorithmic regulation*. Oxford University Press

Yeung, K. (2023). The new public analytics as an emerging paradigm in public sector administration. *Tilburg Law Review*.

Yu, M. C., & Kuncel, N. R. (2020). Pushing the limits for judgmental consistency: Comparing random weighting schemes with expert judgments. *Personnel Assessment and Decisions*, 6(2).

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, *54*, 962–970.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning*, *28*(3), 325–333.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*(4), 661–683.

Zezulka, S., & Genin, K. (2024). From the fair distribution of predictions to the fair distribution of social goods: Evaluating the impact of fair machine learning on long-term unemployment. In *The 2024 ACM conference on fairness, accountability, and transparency* (pp. 1984–2006).

Zschirnt, E., & Ruedin, D. (2016). Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies, 42*(7), 1115–1134.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.