INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

IAB-DISCUSSION PAPER

Articles on labour market issues

06|2026 Distributional Intersectional Fairness in AI-Supported Job Matching: Evidence on Amplification, Trade-Offs, and Residual Risk

Sabrina Mühlbauer, Paula Ziethmann, Enzo Weber

ISSN 2195-2663

Distributional Intersectional Fairness in AI-Supported Job Matching: Evidence on Amplification, Trade-Offs, and Residual Risk

Sabrina Mühlbauer (IAB, Universität Regensburg),
Paula Ziethmann (IAB, Universität Augbsburg),
Enzo Weber(IAB, Universität Regensburg)

Mit der Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

The “IAB-Discussion Paper” is published by the research institute of the German Federal Employment Agency in order to intensify the dialogue with the scientific community. The prompt publication of the latest research results via the internet intends to stimulate criticism and to ensure research quality at an early stage before printing.

Contents

Abstract

Despite widespread recognition of intersectionality in discrimination research, most algorithmic fairness evaluations remain limited to single-attribute group comparisons or individual-level prediction errors. This paper introduces a distributional approach to intersectional fairness that evaluates how machine learning models reshape the allocation of outcomes across intersecting social groups.

Using large-scale administrative labour-market data from public employment services, we analyse an AI-supported job matching system with 141 occupational classes and more than three million observations. We measure intersectional disparities via Jensen–Shannon divergence and compare observed occupational distributions to model-implied allocations.

We show that unconstrained models do not merely reproduce existing labour-market structures but systematically amplify intersectional disparities across occupations. Fairness-aware regularization reduces aggregate disparities but introduces a pronounced trade-off with predictive performance and reallocates distortions across occupations rather than eliminating them. Counterfactual averaging achieves substantial reductions in occupation-level disparities while preserving most predictive performance, may provide a more favourable fairness-performance trade-off in this application setting than directly constraining model outcomes.

To identify localized fairness risks, we propose an occupation-level audit mechanism that flags cases with unusually large distributional shifts. Even under strong fairness constraints, a small number of occupations exhibit substantial residual distortions, indicating that fairness interventions compress rather than remove inequality.

Our results demonstrate that fairness in large-scale decision-support systems must be evaluated at the level of outcome distributions. More broadly, they highlight the limits of purely technical mitigation and the need to combine fairness-aware modelling with explainability, monitoring, and human oversight in real-world deployment.

Zusammenfassung

Obwohl die Bedeutung von Intersektionalität in der Diskriminierungsforschung weithin anerkannt ist, beschränken sich die meisten Evaluationen algorithmischer Fairness weiterhin auf Gruppenvergleiche entlang einzelner Merkmale oder auf Vorhersagefehler auf Individualebene. Dieses Papier entwickelt einen verteilungsbasierten Ansatz zur Untersuchung intersektionaler Fairness, der analysiert, wie maschinelle Lernverfahren die Verteilung von Ergebnissen zwischen sich überschneidenden sozialen Gruppen verändern. Auf Grundlage umfangreicher administrativer Arbeitsmarktdaten der öffentlichen

Arbeitsvermittlung untersuchen wir ein KI-gestütztes Berufsempfehlungssystem mit 141 Berufskategorien und mehr als drei Millionen Beobachtungen. Wir messen intersektionale Unterschiede mithilfe der Jensen-Shannon-Divergenz und vergleichen die beobachteten Verteilungen der Personengruppen über Berufe mit den vom Modell implizierten Allokationen.
Wir zeigen, dass Modelle ohne Fairness-Beschränkungen bestehende Arbeitsmarktstrukturen nicht nur reproduzieren, sondern intersektionale Unterschiede zwischen Berufen systematisch verstärken. Fairness-orientierte Regularisierung reduziert die aggregierten Unterschiede, geht jedoch mit einem ausgeprägten Zielkonflikt zwischen Fairness und Vorhersagegüte einher und verlagert Verzerrungen zwischen Berufen, anstatt sie zu beseitigen. Counterfactual Averaging führt zu einer deutlichen Verringerung berufsspezifischer Unterschiede, während der Großteil der Vorhersagegüte erhalten bleibt. In diesem Anwendungskontext kann dadurch, im Vergleich zu einer direkten Beschränkung der Modellergebnisse, eine günstigere Kombination von Fairness und Vorhersagegüte ermöglicht werden.
Um lokal begrenzte Fairness-Risiken zu identifizieren, schlagen wir einen Prüfmechanismus auf Berufsebene vor, der Fälle mit ungewöhnlich starken Verschiebungen der Verteilungen kennzeichnet. Selbst unter starken Fairness-Beschränkungen weisen einzelne Berufe erhebliche verbleibende Verzerrungen auf. Dies deutet darauf hin, dass Fairness-Maßnahmen Ungleichheiten zwar verringern, aber nicht vollständig beseitigen. Unsere Ergebnisse zeigen, dass Fairness in groß angelegten Entscheidungsunterstützungssystemen auf der Ebene der Ergebnisverteilungen bewertet werden muss. Darüber hinaus verdeutlichen sie die Grenzen rein technischer Ansätze zur Reduzierung von Verzerrungen und die Notwendigkeit Fairness-orientierte Modellierung bei Einsatz in der Praxis mit Erklärbarkeit, dem kontinuierlichem Monitoring und menschlicher Kontrolle zu verbinden.

JEL

J71, J64, C55

Keywords

Algorithmic Decision-Making, Embedded Ethics, Extreme Gradient Boosting, Job Recommendations, Public Sector AI

Danksagung

We would like to thank Anja Mißling-Matthes and Sandra Biermeier from the FOKO department for their valuable support. We are also grateful to colleagues from the DIM department, in particular Joshua Viehrig, for providing the data. We thank Kavita Katdare for her valuable support. Finally, we thank the participants of the CMStatistics 2026 conference for their helpful comments, discussions, and feedback.

1 Introduction

Research on algorithmic fairness has accelerated as it has become unmistakably clear that AI systems can reproduce and amplify patterns of discrimination (Buyl et al., 2022; Poe/El Mestari, 2024). In many technical contexts, such "discrimination" is understood in a narrow, operational sense: machine-learning models are expected to distinguish, rank, or classify, and differential treatment of cases is therefore both inherent and, at times, intended. Yet when claims that "AI discriminates" are interpreted in a social sciences register, the meaning shifts. Here, discrimination refers to the unequal distribution of opportunities or burdens along socially salient lines – such as gender, ethnicity, migration background, or age – shaped by historical inequalities and embedded power relations. The tension between these technical and social-scientific understandings of discrimination and fairness has become a defining challenge in the field (Zollo et al., 2025).

This challenge is particularly pronounced in employment-related AI systems. Recommendation, matching, and profiling models used in labour-market contexts influence access to occupations, training opportunities, and employment trajectories. At the same time, labour markets themselves are already structured by complex patterns of social sorting, occupational concentration, and unequal opportunity. Fairness analysis in such settings therefore faces a dual problem: it must account both for inequalities embedded in the observed data and for distortions introduced through the predictive transformation performed by the model itself.

A growing literature addresses algorithmic fairness in employment and public-sector AI through technical mitigation strategies, governance frameworks, transparency requirements, and human oversight mechanisms (Sharma/Rajput/Srivastav, 2025; Davidson et al., 2026; Broecke, 2023). Yet existing approaches overwhelmingly rely on isolated protected attributes, prediction-level metrics, or aggregate model evaluations. As a consequence, they provide only limited insight into how AI systems reshape outcome distributions across intersecting social groups. This limitation is particularly consequential in labour-market matching, where occupational allocation patterns emerge through the interaction of multiple demographic and socio-economic characteristics.

The importance of intersectionality for understanding discrimination has long been emphasized in the humanities and social sciences. As formulated by Crenshaw (1991), disadvantage does not arise along single and separable dimensions, but through the interaction of multiple social positions. Nevertheless, as recent reviews of fairness research demonstrate, the systematic integration of intersectionality into technical fairness analysis remains limited (Chen et al., 2024; Mehrotra/Pradelski/Vishnoi, 2022). Most empirical studies continue to rely on single-axis subgroup comparisons, despite extensive evidence that disadvantage frequently emerges through intersecting structures of inequality.

Our work is situated within a project that develops an AI-based recommendation system for

public employment services (PES). Previous work has already outlined the ethical, sociotechnical, and discrimination-sensitive challenges associated with such systems (Bauer et al., 2025). We argued there that responsible system development requires not only normative reflection but also empirically grounded approaches capable of identifying and evaluating fairness risks in real-world deployment settings. Related work exists in adjacent contexts. For example, Bied et al. (2023) conduct a fairness analysis for an AI-supported matching project in the French PES, focusing primarily on gender-based disparities. While such studies represent important advances toward fairness-aware system design, they also illustrate a broader limitation of current practice: fairness is usually assessed along single dimensions of discrimination, even though labour-market outcomes are shaped by multiple and interacting social characteristics.

To address this gap, we develop a framework for evaluating intersectional fairness in AI-supported labour-market matching. Our contribution is threefold. Conceptually, we shift the focus from individual prediction errors and single-attribute subgroup comparisons to the distributional consequences of algorithmic recommendations across intersectional social groups. Methodologically, we integrate distributional and predictive fairness measures with occupation-level diagnostics and counterfactual robustness analysis, thereby extending fairness evaluation beyond aggregate metrics. Empirically, we apply this framework to administrative labour-market data covering more than three million observations and 141 occupations, providing evidence on the amplification of intersectional disparities, the fairness-performance trade-off associated with fairness-aware regularization, and the persistence of residual risks despite technical mitigation.

Empirically, we find that unconstrained models systematically amplify intersectional disparities in occupational allocations. Fairness-aware regularization reduces predictive disparities across groups but introduces a pronounced fairness-performance trade-off and leaves residual distributional distortions concentrated in a small number of occupations. Counterfactual averaging (CF) further reduces demographic sensitivity while largely preserving predictive performance, suggesting a more favourable fairness-performance trade-off than direct fairness regularization. Together, these findings highlight both the potential and the limits of technical mitigation: while counterfactual averaging substantially reduces demographic sensitivity at comparatively low utility costs, residual risks persist and require complementary monitoring, explainability, and human oversight.

These findings carry broader implications for fairness research and public-sector AI deployment. First, they suggest that fairness evaluation in employment recommendation systems cannot be reduced to a single metric or a single level of analysis. Second, they indicate that technical mitigation primarily redistributes or compresses fairness risks rather than removing them entirely. Third, they underline the importance of complementing fairness-aware modelling with explainability functions, monitoring mechanisms, human oversight, and governance structures that involve both caseworkers and job seekers as active participants in the recommendation process.

The paper proceeds as follows. Section 2 reviews the literature on algorithmic fairness and

intersectionality. Section 3 introduces the data, modelling framework, and fairness measures. Sections 4 and 5 present the empirical results, including fairness-aware estimation, residual risk analysis, and counterfactual robustness. Section 6 discusses implications for fairness evaluation and AI governance in Public Employment Services. The final section concludes and outlines implications for explainability, participatory design, and human-centered fairness governance in PES.

2 Background and Related Literature

2.1 Existing Strategies on Algorithmic Fairness in Employment Research

The literature on algorithmic fairness in the employment context approaches fairness through technical, organizational, legal, and governance-oriented perspectives. Across these strands, there is broad agreement that AI-supported recruitment and placement systems may reproduce or intensify existing labour-market inequalities. A substantial body of work focuses on the technical origins of algorithmic bias and corresponding mitigation strategies (Sharma/Rajput/Srivastav, 2025). Within the fair machine learning paradigm, discriminatory outcomes are often attributed to biased training data, label constructions, or feature choices, and addressed through techniques such as reweighting, fairness-aware loss functions, or constrained optimization. While these approaches provide important tools for bias mitigation, fairness is typically conceptualized along isolated dimensions, such as single protected attributes or specific bias sources (Agbasiere/Nze-Igwe, 2025). Beyond technical interventions, organizational and governance-oriented research emphasizes that fair outcomes depend not only on model design but also on institutional arrangements, oversight mechanisms, and decision-making practices (Davidson et al., 2026). Studies in the context of public employment services highlight that profiling and matching systems can achieve high predictive performance while simultaneously producing substantial distributive effects across population groups (Desiere/Struyven, 2021; Bach et al., 2023). Rather than relying solely on technical fixes, this literature stresses the importance of transparency, accountability, and deliberation around fairness trade-offs. User-centered research further demonstrates that perceived fairness depends on how algorithmic outputs are interpreted and integrated into decision-making processes (Malin et al., 2025). In public employment services, caseworkers act as intermediaries who contextualize and potentially override algorithmic recommendations, underscoring that fairness emerges through the interaction of system design, organizational routines, and

human judgment (Flügge, 2021; Körtner/Bonoli, 2022). At a more macro level, policy-oriented research and OECD reports situate algorithmic fairness within broader labour market and social policy debates, framing it as a governance challenge that requires institutional steering, transparency, and stakeholder participation (Broecke, 2023; Brioscú et al., 2024). Taken together, existing strategies on algorithmic fairness in employment research provide valuable technical and institutional insights, but they rarely operationalize fairness as a multidimensional and relational phenomenon. Intersectional perspectives remain the exception rather than the rule. As a result, current approaches offer limited tools for capturing how intersecting social positions jointly shape algorithmic outcomes in employment systems. The present study addresses this gap by applying an explicitly intersectional analytical framework to AI-supported labour-market matching.

2.2 Intersectionality as an Analytical Framework for Understanding Inequality

Intersectionality, as formulated by Crenshaw (1991), rests on the insight that discriminatory disadvantage does not occur along a single, neatly isolable axis. Rather, it crystallises through the convergence of multiple dimensions of social differentiation. Crenshaw's intervention addressed a structural blind spot in both antidiscrimination law and feminist theory: by treating categories such as "gender" or "race" as separate and mutually exclusive, institutions and scholarship obscured the ways in which individuals situated at their intersections experience forms of disadvantage that cannot be reduced to either category alone. Her analysis of Black women's lived experiences demonstrated that discriminatory dynamics operate in complex ways – interlocking, overlapping, and mutually reinforcing – and thus reveal inequality as a fundamentally multidimensional phenomenon.

In contemporary intersectionality research more broadly, the central object of inquiry is the emergence and reproduction of social inequalities. Social inequality is understood as the uneven distribution of resources, power, and opportunities for social participation that constrains individuals' life chances over time. Individuals become marginalised when discriminatory practices—whether through formal exclusion, subtle sanctioning, or routine disadvantage – limit their ability to participate fully in social, economic, or political life. Intersectional analysis examines how such marginalisation emerges in particular contexts and how multiple axes of difference jointly shape these processes (Hill Collins/Bildge, 2020).

Which categories of differentiation matter under which conditions is not predetermined (Page/Yip, 2020). Intersectionality theory posits that experiences of discrimination and privilege are relationally shaped by social markers such as race, class, gender, dis/ability, sexuality, or age, while acknowledging that these markers do not operate uniformly across

contexts. The analytical focus therefore lies in understanding how these differentiations interact – how they are embedded in, and sustained by, power relations and institutional arrangements that position groups unevenly within social hierarchies. The aim is not to establish a fixed list of relevant categories but to investigate how multiple dimensions of difference become salient within a given setting. Recognising that AI systems reproduce social inequalities and power structures underscores the importance of transferring this conceptual openness into the examination of algorithmic systems themselves. Because inequality manifests differently across domains—labour markets, education, housing, healthcare – the categories through which it should be analysed inevitably vary. Intersectional analysis thus resists rigid systematisation, insisting instead on attentiveness to context and to the historically contingent nature of social differentiation (Vethman et al., 2025). Making intersecting structures of inequality visible is not merely a descriptive task. By tracing how discriminatory mechanisms arise and interweave, intersectionality points directly to potential sites of intervention within relations of domination. Visibility becomes a precondition for change: only when the layered constitution of inequality is rendered analytically graspable can strategies for its disruption be developed (Bauer et al., 2025). This methodological orientation also explains why intersectionality has never been tied to a single research procedure. As Davis (2008) emphasises, there are no universal rules governing how intersectional research must proceed. The field is characterised by methodological pluralism –from qualitative ethnography to quantitative modelling – by an openness to new analytical tools capable of rendering complex constellations of power intelligible. Intersectionality is therefore best understood not as a fixed methodology, but as an epistemic commitment: to analysing inequality in its relational, situated, and multidimensional form (Hancock, 2007). Our own analysis focuses on the dataset and the model we developed for use as a recommendation system in public employment services. Section 3.1 outlines in detail the composition of our data, the rationale for applying intersectional research to our analytical context, and the way in which we operationalise this perspective in our study.

2.3 Intersectional Analysis of Data and AI

While intersectionality is firmly established in the social sciences and humanities, its systematic integration into data science and AI research remains limited. Only recently have scholars begun to translate intersectional insights into methodological guidance for quantitative and computational contexts. A prominent contribution in this direction is provided by Nielsen et al. (2025), who conceptualize intersectionality not merely as a normative concern but as a methodological orientation that informs variable selection, modelling decisions, and interpretation across the entire analytical pipeline.
A central challenge identified in this literature concerns the context-dependent

identification and operationalization of relevant social categories. Social attributes such as gender, age, or nationality are socially constructed and may interact in non-linear ways, complicating their representation in quantitative models. These difficulties are compounded by data limitations, proxy variables, and reduced statistical power in fine-grained subgroup analyses. Rather than treating these constraints as disqualifying, recent work emphasizes the importance of transparent handling of limitations and empirically grounded subgroup constructions. Our research design follows this logic by combining theory-informed variable selection with large-scale administrative data and by explicitly acknowledging representational gaps in the dataset. In parallel, a growing body of AI fairness research has operationalized intersectionality through subgroup-based performance comparisons, interaction-sensitive statistical testing, or extensions of established fairness concepts (An et al., 2025; Mohammadi et al., 2025; Wang et al., 2024). Seminal work such as Buolamwini/Gebru (2018) Gender Shades study demonstrates how accuracy disparities emerge at the intersection of gender and skin tone, while subsequent contributions propose statistical frameworks to detect interaction effects between multiple sensitive attributes (Suk/Han, 2023). Although these approaches have yielded important insights, they are often tailored to low-dimensional classification tasks and balanced subgroup definitions.

In contrast, the present study addresses a large-scale, highly imbalanced, multiclass matching setting. Rather than focusing on individual-level prediction errors, we adopt a distributional perspective that captures how intersectional groups are structurally positioned across occupational outcomes. This approach allows us to operationalize intersectionality in a way that is compatible with real-world labour-market data and scalable decision-support systems.

While the preceding discussion has focused on methodological and technical challenges, it is important to emphasize that intersectionality is not merely a question of combining variables, but a fundamentally different way of conceptualizing inequality in data-driven systems. Intersectional analysis does not assume that social categories such as gender, age, or nationality exert independent and additive effects. Instead, it highlights that disadvantage emerges through their interaction and that these interactions are shaped by broader structural and institutional dynamics. From this perspective, intersectional group constructions are not neutral statistical artifacts, but representations of socially embedded positions. Their empirical manifestation in data reflects both structural constraints and individual decision-making processes, including self-selection into occupations. As a result, observed disparities cannot be interpreted purely as outcomes of discrimination, but must be understood as the joint result of preferences, opportunities, and institutional sorting mechanisms.

This has direct implications for fairness analysis in machine learning. First, it suggests that fairness cannot be adequately assessed by comparing marginal distributions or isolated group outcomes. Instead, it requires attention to how multiple dimensions of difference jointly structure outcome distributions. Second, it implies that attempts to eliminate all

observed disparities through technical interventions may introduce new distortions if they ignore the underlying socio-economic processes that generate these patterns.
In large-scale applied settings such as labour-market matching, this perspective calls for analytical approaches that remain sensitive to structural inequality while being compatible with high-dimensional data and complex prediction tasks. The distributional approach adopted in this paper follows this logic by focusing on how intersectional groups are positioned across outcomes, rather than on isolated prediction errors. This allows us to capture not only whether disparities exist, but how they are reshaped by the predictive model.

3 Empirical Study: Fairness Metrics, Estimation, and Model Selection

This section introduces the empirical framework used to analyze how occupational outcomes emerge from observed transition patterns and how these patterns interact with fairness considerations in AI-supported job matching. The empirical strategy distinguishes two complementary fairness dimensions. First, we evaluate *distributional fairness* by examining whether the model amplifies or attenuates intersectional concentration patterns across occupational outcomes. This dimension is measured using Jensen-Shannon divergence (JSD). Second, we evaluate *predictive fairness* by examining whether different intersectional groups receive equally accurate occupational recommendations conditional on the true occupational outcome. This dimension is measured using the equal opportunity gap (EOG). The empirical strategy proceeds in four steps. First, we identify status-quo transition patterns into occupations that arise through self-selection and existing labour-market sorting mechanisms and examine their intersectional structure in the observed data. Second, we estimate an unconstrained baseline model and compare its implied occupational distributions to these empirical patterns. Third, we introduce fairness-aware model variants and evaluate their effect on predictive fairness. Fourth, we select a preferred model under a retained-utility criterion.

3.1 Data and Intersectional Setup

The empirical analysis is based on administrative data from the Integrated Employment Biographies (IEB), which provide longitudinal labour-market histories for the German workforce. The dataset is reconstructed to approximate the informational environment of

the Placement and Counseling Information System (VerBIS), thereby aligning the estimation sample with the institutional context of AI-supported counseling. The analytical sample consists of a 10% random sample with complete information, comprising 3,174,204 individuals. The outcome variable is the occupational group after unemployment, measured at the three-digit level of the German Classification of Occupations 2010 (KldB 2010). While the classification distinguishes 144 occupations, we aggregate military occupations into a single category due to small sample sizes, resulting in $K = 141$ occupational classes.

At the core of this analysis is an intersectional perspective. Rather than examining social characteristics in isolation, intersectionality emphasizes that multiple dimensions jointly shape labour-market outcomes. Analyzing gender, age, or nationality separately risks masking important patterns that only emerge through their interaction. Intersectional analysis therefore captures compounded or non-additive effects and provides a more accurate representation of inequality structures. In this study, we focus on three variables with discriminatory potential: gender, nationality, and age. Gender is recorded as a binary variable (male/female), reflecting institutional data constraints. Age is observed with high precision and grouped into meaningful categories capturing life-cycle differences. Nationality serves as a proxy for broader stratification patterns in the absence of direct measures such as race. We distinguish between five groups: German, ten main migration countries, EU, Europe without EU and remaining countries. Individuals belonging to one the ten main migration countries are assigned exclusively to this category and are therefore not represented in one of the other groups. Each individual can belong to exactly one nationality group.

We define intersectional group membership as the joint realization of these three variables. For individual i, let

$$G_i = \left(\text{gender}_i, \text{nationality}_i, \text{age}_i\right) \tag{1}$$

where gender_i, nationality_i, and age_i denote the categorical realizations of the respective attributes. The resulting set of possible combinations is denoted by $\mathcal{G}$, with cardinality $|G| = 60$.

This definition implies that intersectional group membership enters the empirical analysis exclusively as a joint categorical construct. In particular, the empirical design does not include the underlying variables separately once the intersectional grouping is formed. This avoids implicitly imposing additive structures and ensures that all effects are captured at the level of the combined categories. Importantly, these attributes are not used as direct inputs in the main prediction model but are reserved for the fairness analysis. This separation ensures that predictive performance and distributional evaluation remain conceptually distinct.

To characterize baseline inequality, we analyze how individuals from different intersectional groups are distributed across occupations. Let $Y_i \in \{1, \ldots, K\}$ denote the

observed occupational outcome for individual i. The population share of intersectional group $g \in \mathcal{G}$ is denoted by $\pi_g = P(G_i = g)$, and the corresponding share within occupation k by $\pi_{g|k} = P(G_i = g \mid Y_i = k)$. Both quantities are estimated empirically from the data using relative frequencies and provide the benchmark against which model-induced changes are evaluated.

3.2 Prediction Framework

The predictive estimation framework builds directly on Mühlbauer/Weber (2024, 2026). In these studies, job recommendation is formulated as a large-scale multiclass prediction problem and estimated using gradient-boosted decision trees tailored to highly imbalanced occupational outcome structures. We adopt the established estimation framework and extend it through a multi-level intersectional fairness evaluation of the resulting occupational recommendations.

We model occupational outcomes as a multiclass prediction problem. Let X_i denote the feature vector of individual i. For each regularization level $\lambda \geq 0$, the model produces predicted class probabilities

$$\hat{p}_{ik}(\lambda) = P(\hat{Y}_i = k \mid X_i; \lambda), \qquad k = 1, \ldots, K. \tag{2}$$

The case $\lambda = 0$ corresponds to the unconstrained baseline model. Predicted occupational assignments are obtained as

$$\hat{Y}_i(\lambda) = \arg \max_{k \in \{1,\ldots,K\}} \hat{p}_{ik}(\lambda). \tag{3}$$

For each occupation k, we define the predicted distribution of intersectional groups as

$$\hat{\pi}_{g|k}(\lambda) = P(G_i = g \mid \hat{Y}_i(\lambda) = k), \tag{4}$$

estimated empirically via relative frequencies.

3.3 Distributional Fairness and Baseline Model Comparison

To evaluate whether the prediction model reshapes observed intersectional allocation structures, we compare occupational group compositions in the observed data with those implied by model predictions.

We measure deviations between occupational and population compositions using the

Jensen-Shannon divergence (JSD). For each occupation k, define

$$\mathrm{JSD}_k^{\mathrm{true}} = H\left(\frac{\pi_{\cdot|k} + \pi_{\cdot}}{2}\right) - \frac{1}{2}H(\pi_{\cdot|k}) - \frac{1}{2}H(\pi_{\cdot}), \tag{5}$$

where $\pi_{\cdot} = (\pi_g)_{g\in\mathcal{G}}$ and $\pi_{\cdot|k} = (\pi_{g|k})_{g\in\mathcal{G}}$ denote the full distributions over intersectional groups, and $H(p) = -\sum_{g\in\mathcal{G}} p_g \log p_g$ denotes Shannon entropy for a discrete distribution p on $\mathcal{G}$. JSD is used because the empirical object of interest is distributional rather than individual-level: we ask whether occupational group compositions become more or less intersectionally concentrated under model prediction. Unlike asymmetric divergence statistics such as the Kullback–Leibler divergence, JSD is symmetric and bounded, which is important in a highly imbalanced multiclass setting with sparse subgroup support in some occupations. This makes it suitable for comparing observed and predicted occupational compositions across many classes without giving disproportionate weight to very small probability cells. Higher values of $\mathrm{JSD}_k^{\mathrm{true}}$ indicate stronger deviations of an occupation's intersectional composition from the population benchmark. These deviations reflect the status-quo outcome of self-selection and structural sorting processes and therefore constitute the empirical reference point for evaluating model-induced distortions. Appendix Table A2 documents the extent of these disparities across occupations. The results reveal substantial heterogeneity. Highly specialized occupations exhibit particularly strong deviations, indicating concentrated and selective transition patterns, whereas larger and more standardized occupational groups tend to display lower divergence. Overall, these findings highlight that occupational outcomes are shaped by structured transition dynamics across intersecting social dimensions. This underscores why an intersectional framework is essential for a meaningful fairness evaluation.
Using the Jensen-Shannon divergence defined above, the predicted divergence is

$$\mathrm{JSD}_k^{\mathrm{pred}}(\lambda) = \mathrm{JSD}\big(\hat{\pi}_{\cdot|k}(\lambda), \pi_{\cdot}\big), \tag{6}$$

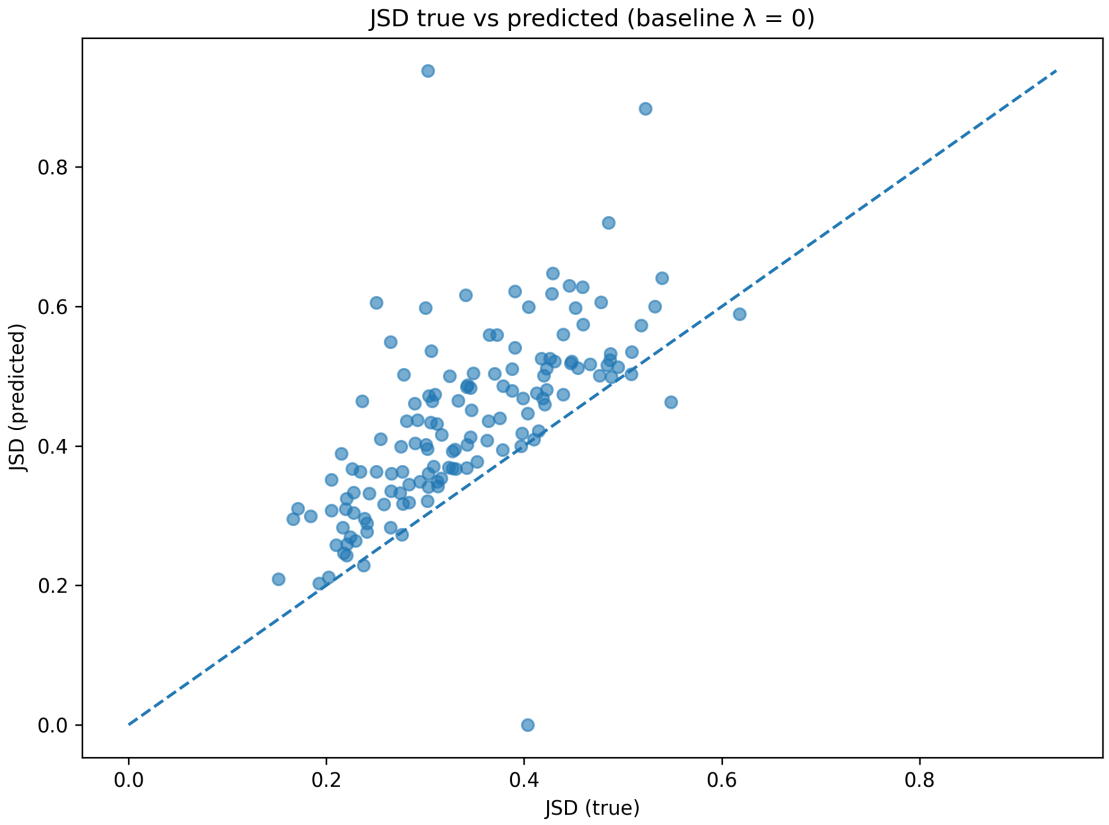
where $\hat{\pi}_{\cdot|k}(\lambda) = (\hat{\pi}_{g|k}(\lambda))_{g\in\mathcal{G}}$. The model-induced shift relative to the empirical benchmark is

$$\Delta\mathrm{JSD}_k(\lambda) = \mathrm{JSD}_k^{\mathrm{pred}}(\lambda) - \mathrm{JSD}_k^{\mathrm{true}}. \tag{7}$$

Positive values indicate that the model amplifies existing distributional disparities, whereas negative values indicate attenuation. Figure 1 shows, for each occupational group, the JSD between the intersectional group distribution within the occupation and the overall population distribution, comparing observed outcomes to model predictions in the unconstrained baseline model ($\lambda = 0$). The 45-degree line indicates perfect reproduction of the empirical distributional structure. Points above this line indicate occupations in which the model amplifies existing intersectional concentration patterns, whereas points below the line indicate attenuation. The figure shows that the unconstrained model does not

simply reproduce the observed intersectional distributions. Instead, it induces systematic shifts across occupations. A non-trivial share of occupations lies above the 45-degree line, indicating prediction-induced amplification of existing disparities, while only a small number of occupations show attenuation. This pattern implies that the baseline model is not distribution-preserving but acts as a selective amplifier of existing labour-market sorting mechanisms.

Figure 1: Occupation-level distributional shifts in the unconstrained baseline model. Values above the 45-degree line indicate amplification of observed intersectional disparities, while values below indicate attenuation.



Source: own calculations. ©IAB

To summarize the most substantively relevant effects, Table 1 reports the ten occupations with the largest positive and negative values of $\Delta \mathrm{JSD}_k(0)$.

The results reveal a strongly asymmetric pattern. The magnitude of amplification substantially exceeds that of attenuation, with the largest positive shifts being considerably larger than the negative ones. This indicates that the unconstrained model systematically increases intersectional concentration in selected occupations. Amplification occurs both in occupations with already high structural concentration and in occupations with only moderate baseline divergence. This suggests that the model captures predictive patterns

Table 1: Occupations with the largest amplification and attenuation of intersectional disparities ($\lambda = 0$)

Occupation	$\mathbf{JSD}_k^{\mathbf{true}}$	$\mathbf{\Delta JSD}_k(0)$
Largest amplifications		
Military	0.303	0.636
Legislators and senior officials of special interest organisations	0.523	0.362
Trading occupations	0.251	0.355
Occupations in publishing and media management	0.300	0.298
Technical occupations in paper-making and -processing and packaging	0.265	0.284
Occupations in artisan craftwork and fine arts	0.341	0.275
Occupations in philology	0.485	0.235
Occupations in mathematics and statistics	0.391	0.231
Sales occupations (retail) selling foodstuffs	0.306	0.230
Occupations in event organisation and management	0.236	0.228
Largest attenuations		
Occupations in economics	0.404	-0.404
Occupations in musical instrument making	0.548	-0.086
Aircraft pilots	0.618	-0.029
Occupations in the production of foodstuffs, confectionery and tobacco products	0.238	-0.009
Occupations in veterinary medicine and non-medical animal health practitioners	0.508	-0.006
Occupations in hotels	0.277	-0.004
Occupations in body care	0.410	-0.000
Occupations in horsekeeping	0.397	0.003
Painters and varnishers, plasterers, occupations in the waterproofing of buildings, preservation of structures and wooden building components	0.414	0.007
Technical occupations in medicine, orthopaedic and rehabilitation	0.202	0.010

Source: own calculations. ©IAB

that reinforce and, in some cases, intensify existing sorting mechanisms. By contrast, attenuation effects are limited and generally small in magnitude. Even in occupations with high baseline divergence, the model rarely produces substantial reductions. The pronounced attenuation observed for economics occupations appears to be an isolated case rather than part of a broader pattern.

Overall, the baseline model acts as a selective amplifier of intersectional disparities. Rather than passively reflecting observed labour-market structures, it actively reshapes them. This distinction is critical: it shows that fairness concerns arise not only from the data-generating process, but also from the predictive transformation applied by the model. The baseline comparison therefore provides the essential reference point for evaluating fairness-aware model adjustments in the subsequent analysis.

3.4 Predictive Fairness, Fairness-Aware Estimation, and Formal Metrics

To evaluate whether the prediction model reshapes observed intersectional allocation structures, we compare occupational group compositions in the observed data with those implied by predictions. The baseline analysis has shown that the unconstrained model does not merely reproduce observed labour-market structures but actively reshapes them, often amplifying intersectional disparities. This raises the central question of whether such distortions can be mitigated through fairness-aware model design, and at what cost in terms of predictive performance.

We estimate a sequence of models indexed by $\lambda \in \Lambda \subset \mathbb{R}_{\geq 0}$. Model training is based on

$$\mathcal{L}(\lambda) = \mathcal{L}_{\text{pred}} + \lambda \mathcal{P}_{\text{fair}}, \tag{8}$$

where $\mathcal{L}_{\text{pred}}$ denotes the class-weighted multiclass log loss and $\mathcal{P}_{\text{fair}}$ penalizes disparities across intersectional groups. The parameter λ governs the trade-off between predictive accuracy and fairness.

The equal opportunity gap $\text{EOG}(\lambda)$ captures predictive fairness. It measures the extent to which individuals from different intersectional groups have unequal chances of being correctly assigned to their true occupation by the model, conditional on that occupation being the correct outcome. For occupation k and group g, let

$$\text{TPR}_{k,g}(\lambda) = P\big(\hat{Y}_i(\lambda) = k \mid Y_i = k, G_i = g\big) \tag{9}$$

denote the group-specific true positive rate. The occupation-specific equal opportunity gap is then

$$\text{EOG}_k(\lambda) = \max_{g \in \mathcal{G}_k} \text{TPR}_{k,g}(\lambda) - \min_{g \in \mathcal{G}_k} \text{TPR}_{k,g}(\lambda), \tag{10}$$

where $\mathcal{G}_k \subseteq \mathcal{G}$ denotes the set of intersectional groups with sufficient support in occupation k. We aggregate this measure as

$$\text{EOG}(\lambda) = \frac{1}{|K^*|} \sum_{k \in K^*} \text{EOGGap}_k(\lambda), \tag{11}$$

where K^* denotes the set of sufficiently supported occupations.

Equal opportunity is selected because the central policy concern in Public Employment Services is not equal occupational assignment rates across demographic groups, but equal predictive access to correct recommendations conditional on the true occupational outcome. Demographic parity would be difficult to justify in this setting because occupational transitions are strongly qualification-dependent and empirically heterogeneous. Equalized odds is substantially harder to interpret in a high-dimensional

multiclass setting with many occupation-by-group combinations and sparse subgroup support. Calibration, in turn, addresses the reliability of predicted probabilities rather than equality in successful recommendation performance. Equal opportunity therefore provides a transparent measure of predictive fairness that directly captures disparities in correct occupational recommendations across intersectional groups.
In the context of AI-supported job matching, this quantity is directly linked to substantive fairness. A large gap implies that some groups are systematically less likely to receive correct recommendations for occupations that are in fact suitable for them. This can be interpreted as unequal access to accurate job matching, which may translate into differential labour-market opportunities. For example, a value of $\mathrm{EOG} = 0.74$ implies that, on average, the difference between the most and least advantaged groups within an occupation is 74 percentage points. Conversely, a reduction of EOG to values around 0.19 implies a much more balanced distribution of predictive accuracy across groups, although this improvement may come at significant cost in predictive performance.
Importantly, $\Delta\mathrm{JSD}_k(\lambda)$ and $\mathrm{EOG}(\lambda)$ capture conceptually distinct fairness dimensions and need not move in parallel. JSD evaluates whether predicted occupational assignment structures become more or less intersectionally concentrated relative to observed labour-market distributions. EOG evaluates whether predictive performance differs systematically across intersectional groups. A model may therefore improve predictive fairness while still increasing distributional concentration, or it may reduce distributional sensitivity while leaving performance disparities across groups largely unchanged. Evaluating both dimensions jointly is necessary because AI-supported occupational recommendation systems can affect both who is predicted accurately and how groups are distributed across predicted occupational outcomes. Predictive performance is evaluated using complementary metrics reflecting the imbalanced multiclass nature of the problem. Log loss captures probabilistic accuracy. Standard accuracy is reported but is not central due to class imbalance. Balanced accuracy averages recall across occupations and is therefore more appropriate for evaluating class-sensitive performance. Top-k accuracy, reported for $k = 3, 5, 10$, is particularly relevant because the system provides ranked occupation recommendations rather than deterministic assignments. Macro one-vs-rest PR-AUC captures class discrimination under imbalance, while the Matthews correlation coefficient provides a balanced overall summary.
Table 2 summarizes the joint performance and fairness outcomes across λ.

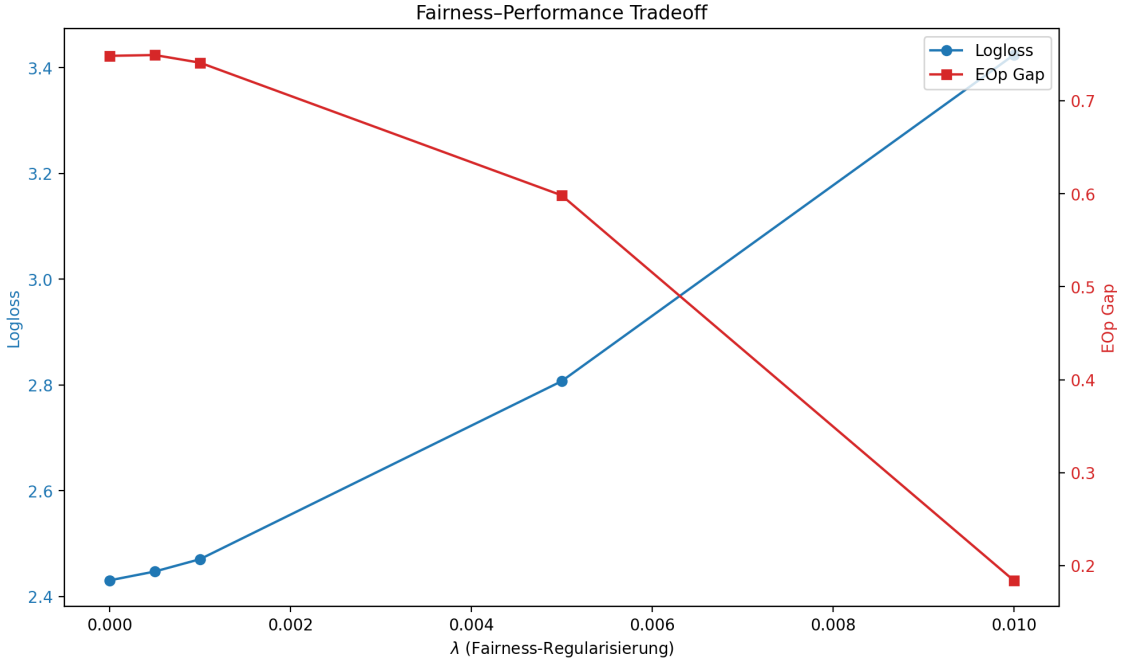
Figure 2 visualizes the core trade-off between log loss and $\mathrm{EOG}(\lambda)$. The results reveal a clear and systematic trade-off between predictive performance and predictive fairness. For very small values of λ ($\lambda \leq 0.001$), predictive performance remains largely stable across all metrics, while fairness improvements are negligible. For intermediate regularization ($\lambda = 0.005$), the equal opportunity gap decreases from 0.74 to 0.60, but balanced accuracy, top-k accuracy, and PR-AUC decline noticeably. For strong regularization ($\lambda = 0.010$), predictive fairness improves substantially, with the equal opportunity gap reduced to 0.19,

Table 2: Performance–fairness trade-off across regularization levels

	Performance							Fairness
λ	LogLoss	Acc	BalAcc	Top-3	Top-5	PR-AUC	MCC	EOG
0.000	2.43	0.524	0.312	0.641	0.722	0.284	0.271	0.747
0.0005	2.45	0.522	0.310	0.639	0.720	0.282	0.269	0.744
0.001	2.47	0.519	0.307	0.636	0.718	0.279	0.265	0.739
0.005	2.81	0.472	0.270	0.590	0.675	0.245	0.231	0.598
0.010	3.45	0.381	0.198	0.480	0.560	0.180	0.162	0.184

Source: own calculations. ©IAB

Figure 2: Fairness–performance trade-off: log loss and equal opportunity gap as functions of λ.



Source: own calculations. ©IAB

but predictive performance deteriorates to a degree that is unlikely to be operationally acceptable in a recommendation setting.

Two central insights emerge. First, predictive fairness improvements are costly in this setting: reducing disparities in correct recommendations across intersectional groups requires the model to deviate from predictive optimality. Second, the trade-off is nonlinear. Small increases in λ preserve performance but have little fairness effect, whereas substantial fairness gains require disproportionately large sacrifices in predictive quality. Top-k accuracy makes this trade-off especially tangible for the application context: at high levels of regularization, even the recommendation set no longer provides reliable guidance.

3.5 Model Selection under Retained Utility Constraints

The previous section established a clear trade-off between predictive fairness and predictive performance. Model selection in this context cannot be reduced to a single-objective optimization problem. Instead, it requires an explicit criterion that balances fairness improvements against the preservation of predictive usefulness.
We formalize this trade-off using retained utility relative to the unconstrained baseline model:

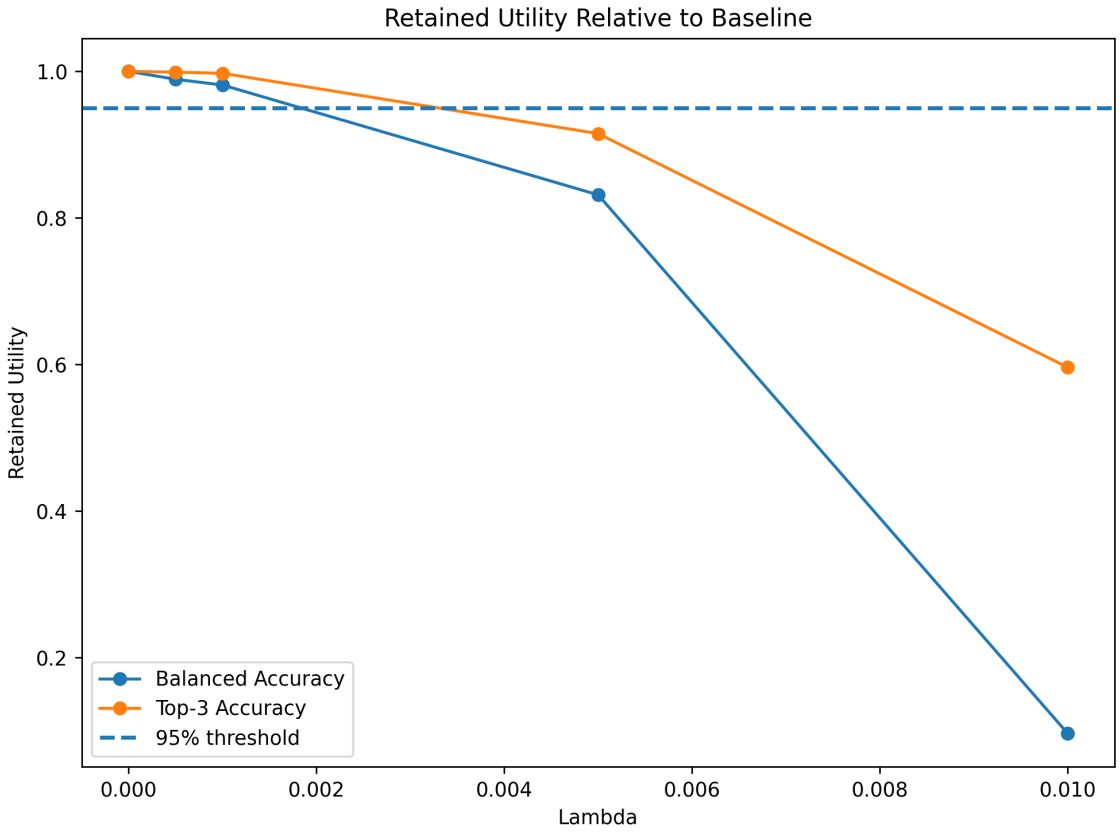
$$\mathrm{RU}(\lambda) = \min\left\{\frac{\mathrm{BalAcc}(\lambda)}{\mathrm{BalAcc}(0)}, \frac{\mathrm{Top3}(\lambda)}{\mathrm{Top3}(0)}\right\}. \tag{12}$$

Here, $\mathrm{BalAcc}(\lambda)$ denotes balanced accuracy and $\mathrm{Top3}(\lambda)$ denotes top-3 accuracy for the model with regularization parameter λ. The baseline values $\mathrm{BalAcc}(0)$ and $\mathrm{Top3}(0)$ serve as normalization benchmarks.
The use of top-3 accuracy is particularly important in the present application. $\mathrm{Top3}(\lambda)$ measures whether the true occupation is contained within the three highest-ranked model recommendations. In an AI-supported job matching system, recommendations are not used as deterministic assignments but as ranked suggestions for caseworkers and job seekers. Preserving top-3 performance therefore ensures that the system continues to provide relevant and actionable guidance.
A model is considered technically admissible if

$$\mathrm{RU}(\lambda) \geq 0.95. \tag{13}$$

The 95% threshold is motivated by the operational context. In a high-stakes application such as public employment services, predictive performance is directly linked to labour-market outcomes. The primary objective of the system is to support effective job matching and thereby contribute to reducing unemployment duration. A model that substantially degrades predictive quality, even if it achieves strong fairness improvements, would undermine this objective.
Figure 3 shows the retained utility profiles for balanced accuracy and top-3 accuracy across λ. The figure reveals that both performance measures remain close to baseline levels for small values of λ, but diverge sharply as regularization increases. Balanced accuracy deteriorates more rapidly than top-3 accuracy, indicating that class-level discrimination is more sensitive to fairness constraints than recommendation quality. For $\lambda = 0.001$, both measures remain above the 95% threshold. At $\lambda = 0.005$, top-3 accuracy is close to the threshold, while balanced accuracy already falls below it. For $\lambda = 0.010$, both measures collapse substantially.
Combining these results with the fairness analysis yields a clear selection outcome. While predictive fairness improves monotonically as λ increases, only small values of λ satisfy the

Figure 3: Retained utility relative to the baseline model. The dashed line indicates the 95% admissibility threshold.



Source: ©IAB

retained utility constraint. Among these, $\lambda = 0.001$ provides the strongest fairness improvement without violating the performance requirement. We therefore select $\lambda = 0.001$ as the technically preferred model. By contrast, $\lambda = 0.010$ can be interpreted as a stress-test scenario: it demonstrates the upper bound of fairness improvements achievable through regularization, but at a level of predictive degradation that renders the model unsuitable for deployment.

This result highlights a central tension in fairness-aware AI systems. The objective is not to achieve maximal fairness in isolation, but to improve fairness within the constraints of practical usefulness. In the present application, the ultimate goal is to identify suitable occupations for job seekers and thereby improve labour-market outcomes. A perfectly fair model that fails to provide accurate recommendations would be counterproductive. Moreover, observed disparities partly reflect endogenous labour-market sorting, including self-selection by individuals. Job seekers do not enter occupations randomly; their choices are shaped by preferences, qualifications, and constraints. Eliminating all observable disparities without accounting for these mechanisms may therefore introduce a different form of distortion. Not all observed differences are necessarily unjust, and not all reductions in disparity improve real-world outcomes.

4 Residual Distributional Risk, Explainability, and Governance

The preceding section identified $\lambda = 0.001$ as the technically preferred model under the retained-utility criterion and showed that stronger regularization can improve aggregate predictive fairness only at substantial performance cost. This section returns to the distributional fairness perspective and asks where residual occupation-level distortions remain concentrated after regularization. The objective is not only to quantify remaining risks, but also to translate them into audit, explainability, and governance mechanisms for Public Employment Services.

4.1 Residual Distributional Risk and Flagged Occupations

While the equal opportunity gap summarizes predictive fairness at the model level, the fairness flag identifies occupations in which the model induces unusually large distributional distortions relative to the empirical benchmark.
Formally, for each λ, an occupation k is flagged if

$$\Delta\mathrm{JSD}_k(\lambda) > \tau(\lambda), \qquad \tau(\lambda) = \mu_\Delta(\lambda) + \sigma_\Delta(\lambda), \tag{14}$$

where $\mu_\Delta(\lambda)$ and $\sigma_\Delta(\lambda)$ denote the mean and standard deviation of $\Delta\mathrm{JSD}_k(\lambda)$ across sufficiently supported occupations.
The fairness flag is not a legal or normative classification of discrimination. Rather, it constitutes an operational audit signal that identifies occupations in which the model induces unusually large distributional distortions relative to the empirical benchmark. As such, it provides a systematic link between aggregate fairness evaluation and localized diagnostic analysis, enabling targeted explainability, monitoring, and governance interventions.
Importantly, the flag is defined relative to the distribution of occupation-level distortions within each model specification. The threshold depends on both the mean distortion $\mu_\Delta(\lambda)$ and the dispersion of distortions $\sigma_\Delta(\lambda)$. A larger standard deviation increases the threshold and therefore makes it more difficult for an occupation to be flagged, whereas a smaller standard deviation lowers the threshold and makes comparatively smaller residual deviations more visible. The flag should therefore not be interpreted as a fixed absolute measure of unfairness across different values of λ. Instead, it identifies occupations whose distributional distortions are unusually large relative to the overall distortion profile of the respective model.

This distinction is particularly important for the strongly regularized model with $\lambda = 0.010$. In this specification, the threshold is substantially higher than in the weakly regularized models. Occupations that remain flagged under this condition therefore exceed a much stricter benchmark. This does not imply that these occupations are necessarily treated less fairly than under the unconstrained baseline in every respect. Rather, it indicates that, even after strong regularization has improved aggregate predictive fairness, these occupations remain exceptional residual cases in distributional terms. In substantive terms, fairness-aware regularization appears to reduce broad-based predictive disparities while leaving a smaller set of occupations with unusually persistent intersectional concentration patterns. These occupations should therefore be interpreted as priority cases for further audit, explainability, and human oversight rather than as direct evidence of legal discrimination. Table 3 reports the number of flagged occupations for each value of λ.

Table 3: Flagged occupations by regularization level

λ	Number Flagged occupations	Threshold	Mean flagged $\Delta\mathrm{JSD}$	Max $\Delta\mathrm{JSD}$
0.0000	20	0.119	0.186	0.789
0.0005	20	0.124	0.197	0.767
0.0010	21	0.126	0.196	0.548
0.0050	21	0.177	0.253	0.808
0.0100	5	0.580	0.637	0.905

Source: own calculations. ©IAB

Table 3 shows that the number of flagged occupations must be interpreted together with the threshold and the severity of the flagged distortions. For weak and intermediate regularization levels, the threshold remains relatively low, between 0.119 and 0.177, and around twenty occupations are flagged. For $\lambda = 0.010$, by contrast, the threshold increases sharply to 0.580. This substantially higher threshold means that occupations must exhibit much larger residual distributional distortions to be flagged at this level of regularization. The decline from around twenty flagged occupations to five therefore does not mean that all other distributional risks disappear mechanically; rather, it indicates that only a small set of occupations remains sufficiently extreme to exceed a much stricter benchmark.
At the same time, the mean flagged $\Delta\mathrm{JSD}$ rises to 0.637 and the maximum $\Delta\mathrm{JSD}$ reaches 0.905. This pattern indicates that strong regularization compresses broad-based distortions into fewer but more severe residual cases. In other words, stronger fairness regularization improves aggregate predictive fairness and reduces the number of occupations that stand out relative to the model-specific threshold, but the occupations that still exceed this threshold exhibit particularly pronounced distributional shifts. These flagged occupations should therefore be interpreted as structurally persistent residual risks rather than as evidence that the overall model has become less fair.
The results show that stronger regularization does not monotonically eliminate flagged occupations. For small and intermediate values of λ, the number of flagged occupations

remains broadly stable. Only for the strongest specification, $\lambda = 0.010$, does the number of flagged occupations fall substantially. However, the remaining flagged occupations exhibit large distortions. This indicates that regularization compresses fairness risks into fewer, but more extreme, residual cases.

The model with $\lambda = 0.010$ achieves the lowest aggregate equal opportunity gap and can therefore be interpreted as the fairest model in predictive terms. Nevertheless, five occupations remain flagged from a distributional perspective. Table 4 lists these occupations. These results are substantively important. Even under the strongest fairness

Table 4: Flagged occupations in the aggregate-fairest model ($\lambda = 0.010$)

Code	Occupation	N_k^{true}	N_k^{pred}	$\Delta\mathrm{JSD}_k$	Main overrepresented group
633	Gastronomy occupations	15,976	343	0.736	Female, Remaining, 25–34
293	Cooking occupations	14,129	403	0.670	Male, Remaining, 55–64
814	Human medicine and dentistry	4,259	205	0.541	Female, Remaining, 25–34
621	Retail trade without product specialisation	29,441	34,806	0.643	Female, Germany, 15–24
921	Advertising and marketing	10,680	54	0.593	Female, Remaining, 25–34

Source: own calculations. ©IAB

regularization, residual distributional distortions remain concentrated in specific occupations. The pattern is not random. The flagged occupations include hospitality, cooking, retail, medicine, and advertising. These are occupational fields in which labour-market sorting is often strongly structured by gender, nationality, age, sectoral experience, and employment histories. Such variables may enter the model as legitimate predictors, but they can also operate as proxies for intersectional group membership.

The relationship between true and predicted class counts is also informative. For some occupations, such as advertising and marketing, human medicine and dentistry, gastronomy, and cooking, the model predicts far fewer cases than are observed in the data. This suggests that strong regularization narrows the model's assignment behavior and concentrates predictions on a small subset of profiles. For retail, by contrast, predicted counts exceed observed counts, indicating that the model reallocates individuals into a broad service occupation. Both patterns can generate large changes in intersectional composition.

The key implication is that fairness-aware regularization does not remove structural inequality; it changes its location and form. At $\lambda = 0.010$, predictive fairness improves substantially, but some occupations remain strongly distorted in distributional terms. This matters because the policy objective is not to produce a formally fair model in isolation. The purpose of the system is to support suitable job matching and thereby contribute to reducing unemployment. A model that is highly regularized but no longer provides reliable recommendations would fail this objective.

4.2 Explainability, Human Oversight, and Jobseeker Participation

The fairness flag is intended to support sociotechnical governance rather than replace professional or individual judgment. In Public Employment Services, algorithmic recommendations are embedded in a decision process that involves not only caseworkers, but also jobseekers themselves. Caseworkers interpret recommendations in light of counselling goals, institutional rules, local labour-market conditions, and available support instruments. Jobseekers, in turn, bring preferences, constraints, qualifications, care responsibilities, mobility limitations, health conditions, and occupational aspirations into the matching process. A fairness-sensitive recommendation system must therefore support both professional interpretation and meaningful participation by jobseekers.

Flagged occupations can serve as targeted triggers for explainability and oversight mechanisms. First, they can be prioritized for local explanation methods that identify which features drive a recommendation and whether employment history, qualifications, sectoral experience, regional labour-market information, or other variables act as proxies for intersectional group membership. Second, flagged recommendations can be accompanied by explanation cards that distinguish qualification-related drivers from potentially sensitive or proxy-related patterns. Third, uncertainty indicators can inform caseworkers and jobseekers when the model produces recommendations in occupational fields with known residual distributional distortions. Fourth, contrastive explanations can show which factors would make alternative occupations more likely, thereby supporting counselling conversations rather than presenting recommendations as fixed assignments.

These mechanisms are important because fairness risks do not arise only from the internal model structure, but also from how model outputs are interpreted and used. Without explanation, a flagged recommendation may be accepted mechanically, even if it reflects residual occupational concentration rather than an individually appropriate match. With appropriate explanation, the same recommendation can become a starting point for critical reflection: caseworkers can assess whether the recommendation is driven by relevant qualifications, jobseekers can contest or contextualize the suggestion, and both parties can consider alternative occupational pathways. For jobseekers, explainability is particularly important because occupational recommendations may shape perceived opportunities and future search behavior. If a system repeatedly recommends occupations that reflect historical sorting patterns, it may narrow rather than expand labour-market opportunities. Explanation interfaces should therefore not only justify why an occupation was recommended, but also support agency by making visible which alternative occupations are plausible, which skills or qualifications would increase access to them, and whether recommendations are influenced by structurally patterned employment histories.

From a governance perspective, the fairness flag therefore provides an operational bridge between statistical fairness diagnostics and practical decision support. Rather than

auditing all occupations equally, governance resources can be focused on occupations where the model produces the strongest residual distortions. These occupations can be monitored over time, reviewed with domain experts, and incorporated into interface-level safeguards. In this sense, the fairness flag complements the aggregate equal opportunity gap: the equal opportunity gap shows whether predictive fairness improves on average, while the fairness flag shows where residual distributional risks remain. Together, they demonstrate that technical mitigation must be combined with explainability, monitoring, jobseeker participation, and human judgment.

5 Counterfactual Averaging and Distributional Fairness

To further evaluate the robustness of the proposed fairness-aware estimation framework, we additionally implement a counterfactual averaging prediction strategy. In this approach, predictions are repeatedly generated under alternative intersectional identity assignments and subsequently averaged before the final occupation recommendation is produced. Importantly, the analysis remains fully intersectional. The counterfactual procedure operates on the complete intersectional representation, which jointly captures combinations of gender, nationality, and age group. The method therefore evaluates whether occupational predictions remain stable across alternative intersectional group assignments.

Conceptually, counterfactual averaging differs fundamentally from fairness regularization. Whereas the regularized model constrains the estimation process directly through the penalty parameter λ, counterfactual averaging modifies the otherwise unconstrained prediction mechanism ex post by averaging predicted probabilities across alternative intersectional identities. Counterfactual averaging therefore primarily targets residual distributional sensitivity rather than predictive parity. Its effects are evaluated mainly through the occupation-level divergence framework, while changes in predictive performance metrics are interpreted as secondary robustness outcomes.

Residual demographic sensitivity captures the remaining dependence of occupational recommendations on intersectional identity assignments after holding all qualification- and employment-related characteristics fixed.

5.1 Aggregate Fairness Effects

Table 5 compares the standard prediction rule with counterfactual averaging across regularization levels.

Table 5: Comparison of flagged occupations under the standard prediction rule and counterfactual averaging (CF).

λ	Flagged (Base)	Flagged (CF)	Mean ΔJSD (Base)	Mean ΔJSD (CF)
0.0000	20	22	0.186	0.134
0.0005	20	24	0.197	0.133
0.0010	21	22	0.196	0.138
0.0050	22	18	0.253	0.186
0.0100	5	2	0.637	0.282

Source: own calculations. ©IAB

The results reveal a differentiated relationship between fairness regularization and counterfactual averaging. For small values of λ, counterfactual averaging does not systematically reduce the number of flagged occupations and occasionally increases it slightly. However, even in these settings, the average severity of the flagged distortions declines noticeably. This indicates that counterfactual averaging primarily weakens the intensity of occupation-level concentration patterns rather than uniformly eliminating them.

The strongest effect emerges for $\lambda = 0.010$. Under the standard prediction rule, five occupations remain flagged, with an average residual divergence of 0.637. Counterfactual averaging reduces the number of flagged occupations to two and lowers the mean residual divergence to 0.282. Substantively, this implies that the remaining prediction distortions become less than half as large after counterfactual smoothing.

This pattern suggests that fairness regularization and counterfactual averaging address different components of the fairness problem. The regularization term primarily suppresses the formation of strongly group-dependent prediction structures during estimation, whereas counterfactual averaging reduces the sensitivity of final predictions to specific intersectional identity assignments. The two mechanisms therefore appear to reinforce each other particularly in highly regularized models.

Counterfactual averaging successfully reduces measured distributional disparities by lowering the dependence of occupational on intersectional identity assignments. However, these reductions should not automatically be interpreted as the complete removal of underlying structural inequalities. Because the approach modifies the prediction mechanism itself through probability averaging across identity perturbations, part of the observed fairness gain reflects reduced demographic sensitivity in prediction space rather than a full disappearance of labour-market inequalities embedded in the underlying data-generating process.

5.2 Predicitve Performance under Counterfactual Averaging

Table 6 compares selected predictive performance metrics between the standard prediction rule and counterfactual averaging for the baseline setting ($\lambda = 0$).

Table 6: Predictive performance comparison between the standard prediction rule and counterfactual averaging for $\lambda = 0$.

Metric	Baseline	Counterfactual Averaging
Balanced Accuracy	0.274	0.262
Top-3 Accuracy	0.603	0.581
MCC	0.360	0.347

Source: own calculations. ©IAB

The performance results indicate that counterfactual averaging slightly reduces predictive accuracy relative to the standard prediction rule. This pattern is theoretically plausible because the averaging procedure intentionally smooths prediction probabilities across alternative intersectional identity assignments. As a consequence, predictions become less specialized toward specific demographic profiles, which can moderately reduce classification precision and ranking performance.

At the same time, the observed utility reductions remain comparatively limited relative to the substantial fairness improvements documented in the occupation-level divergence analysis. In particular, counterfactual averaging substantially lowers residual demographic sensitivity, meaning that occupational recommendations become less dependent on specific combinations of gender, nationality, and age group while preserving most of the overall predictive structure of the model.

The two mitigation strategies operate through fundamentally different mechanisms. Fairness-aware regularization directly constrains model outcomes and therefore reduces disparities irrespective of their source. Counterfactual averaging, by contrast, targets only the contribution of protected characteristics while leaving labour-market-relevant predictors unchanged. As a result, some distributional disparities remain even after counterfactual adjustment. However, these residual differences are more likely to reflect variation in qualifications, employment histories, or other job-relevant characteristics rather than the direct influence of protected attributes. This distinction helps explain why counterfactual averaging achieves substantial reductions in distributional disparity while maintaining predictive performance.

For PES, this trade-off may be particularly attractive because recommendation systems operate in institutionally sensitive environments where procedural consistency and fairness are central operational requirements. Counterfactual averaging therefore appears to provide an additional robustness layer that can complement fairness-aware estimation while maintaining comparatively strong predictive utility.

5.3 Occupation-Level Comparison for the Strongest Regularization Setting

Table 7 reports the occupations that remain flagged under the strongest regularization setting ($\lambda = 0.010$) and compares the standard prediction rule with counterfactual averaging. The occupation-level results clarify the magnitude of these effects. For trading

Table 7: Flagged occupations under the standard prediction rule and counterfactual averaging for $\lambda = 0.010$.

Occupation	JSD_{true}	JSD_{base}	JSD_{cf}	ΔCF vs. Base
Trading occupations	0.251	0.888	0.514	-0.374
Occupations in mechatronics, automation and control engineering	0.404	0.932	0.650	-0.282

occupations, the predicted intersectional distribution under the standard prediction rule deviates very strongly from the observed distribution, with a JSD of 0.888 relative to an observed benchmark of 0.251. Counterfactual averaging reduces this divergence by 0.374, corresponding to a reduction of approximately 42%. A similar pattern emerges for occupations in mechatronics, automation and control engineering, where the divergence declines from 0.932 to 0.650, implying a reduction of roughly 30%. In substantive terms, the results suggest that part of the residual unfairness in highly regularized models stems from prediction sensitivity to specific intersectional identities rather than exclusively from broader structural imbalances in the training data. At the same time, the remaining divergence levels after counterfactual averaging remain substantially above the observed benchmark distributions. This is important because it indicates that even aggressive post-processing interventions cannot fully eliminate occupation-level concentration patterns. The persistence of residual divergence therefore supports the interpretation that some disparities reflect underlying labour-market structures that are already embedded in the observed data-generating process.

Importantly, the persistence of some distributional disparities under counterfactual averaging should not automatically be interpreted as a fairness failure. Because the intervention neutralizes protected attributes while preserving labour-market-relevant information, remaining disparities may partly reflect differences in qualifications, employment histories, or occupational specialization rather than demographic characteristics themselves.

The counterfactual results additionally highlight several properties that are particularly relevant for operational deployment in PES. Counterfactual averaging reduces the sensitivity of occupational recommendations to demographic characteristics while preserving the broader predictive structure of the model. In practical terms, this means that occupational recommendations become more robust against demographic perturbations and less dependent on specific combinations of gender, nationality, and age.

This property is especially valuable in PES because recommendation systems are typically

used in highly heterogeneous populations characterized by diverse educational, social, and demographic backgrounds. Counterfactual averaging therefore offers an additional safeguard against the emergence of strongly group-dependent recommendation patterns and may improve the perceived procedural fairness and consistency of algorithmic support systems. The empirical results show the effects when counterfactual averaging is combined with fairness-aware estimation. Whereas regularization reduces the formation of structurally unfair prediction patterns during model training, counterfactual averaging attenuates residual demographic sensitivity at the prediction stage. The two mechanisms therefore appear to complement each other by operating at different stages of the modeling pipeline. At the same time, the moderate reduction in predictive performance observed under counterfactual averaging is theoretically plausible. By construction, the procedure smooths prediction probabilities across alternative demographic constellations and thereby partially suppresses information that may still carry predictive signal within the observed data-generating process. As a result, prediction outputs become less specialized toward particular intersectional profiles, which can slightly reduce classification precision and ranking performance. Importantly, however, the results indicate that these utility reductions remain comparatively limited relative to the observed fairness gains.

Comparing both approaches reveals an important difference. Fairness-aware regularization reduces disparities primarily through stronger constraints on the prediction process, which eventually leads to substantial declines in predictive performance. Counterfactual averaging instead removes demographic sensitivity while preserving the informational contribution of non-protected predictors. Consequently, it achieves a more favourable fairness-performance trade-off, reducing occupation-level disparities without requiring the extensive predictive sacrifices observed under strong regularization.

For practical deployment in PES, counterfactual averaging would nevertheless require additional supporting mechanisms. In particular, XAI components remain important to ensure transparency and accountability of recommendations. Caseworkers and jobseekers must be able to understand how recommendations are generated, which qualification-related characteristics remain most influential, and how fairness interventions affect prediction probabilities. Overall, the findings suggest that counterfactual averaging represents a promising complementary fairness mechanism for PES. The approach appears particularly valuable as an additional robustness layer that reduces residual demographic sensitivity while maintaining comparatively strong predictive utility.

6 Discussion: Integrating Fairness Dimensions in AI-Supported Job Matching

The empirical results highlight that fairness in AI-supported job matching is inherently multidimensional and cannot be adequately captured by a single metric or analytical perspective. Our findings show that distributional fairness, predictive fairness, and operational fairness risk are related but distinct properties of the modeling pipeline. Considering these dimensions jointly substantially changes how fairness interventions must be interpreted and evaluated in labour-market recommendation systems.

6.1 Distributional and Predictive Fairness Do Not Coincide

A central finding of the analysis is that distributional fairness and predictive fairness capture different aspects of model behaviour and should therefore not be conflated. The Jensen–Shannon divergence framework evaluates how occupational outcome distributions across intersectional groups are reshaped by the predictive model relative to the empirical benchmark. By contrast, the equal opportunity gap evaluates whether predictive performance is distributed evenly across groups conditional on the true occupational outcome. The empirical results demonstrate that these fairness dimensions do not necessarily move together. The baseline model exhibits substantial amplification of occupational concentration patterns in distributional terms while simultaneously maintaining comparatively strong predictive performance. Conversely, strong fairness regularization substantially reduces predictive disparities as measured by the equal opportunity gap but does not fully eliminate occupation-level distributional distortions. This distinction is substantively important for employment-related AI systems. A model may provide relatively balanced predictive accuracy across groups while still reallocating occupational opportunities in ways that increase intersectional concentration. Similarly, reductions in aggregate predictive disparities do not automatically imply preservation of observed occupational compositions. The results therefore suggest that fairness evaluation in labour-market recommendation systems requires a structured multi-level framework rather than reliance on isolated fairness metrics.

6.2 Technical Mitigation Compresses Rather Than Eliminates Fairness Risk

The results further indicate that technical fairness interventions primarily reshape the location and form of fairness risk rather than removing it entirely. Fairness-aware regularization reduces predictive disparities across intersectional groups but introduces a pronounced and nonlinear trade-off with predictive performance. Small regularization levels preserve most predictive utility while producing limited fairness improvements, whereas substantial reductions in predictive disparities require increasingly large sacrifices in recommendation quality. The occupation-level fairness flag reveals an additional layer of complexity. Aggregate fairness improvements do not imply uniform reductions in residual risk across occupations. Instead, stronger regularization tends to compress fairness risk into a smaller number of occupations characterized by unusually large residual distortions. Importantly, the threshold defining flagged occupations depends on the distribution of occupational shifts within a given specification. Because the threshold is defined relative to the mean and standard deviation of $\Delta\mathrm{JSD}$, cross-model comparisons must be interpreted carefully. A smaller number of flagged occupations under strong regularization therefore does not necessarily imply that all occupations become uniformly fairer. Rather, it indicates that fairness risk becomes concentrated in fewer but more pronounced residual cases relative to the distributional benchmark of that specification. The counterfactual robustness analysis reinforces this interpretation. Counterfactual averaging reduces residual demographic sensitivity, particularly in highly regularized models, indicating that some residual distortions remain sensitive to alternative intersectional identity assignments. At the same time, substantial residual divergence persists even after counterfactual smoothing. This suggests that fairness-aware estimation and counterfactual averaging address different components of the fairness problem. Regularization constrains the formation of group-dependent prediction structures during estimation, whereas counterfactual averaging attenuates demographic sensitivity at the prediction stage. Neither mechanism fully removes the influence of underlying labour-market sorting structures embedded in the observed data.

This observation also helps explain why the two mitigation strategies exhibit markedly different fairness-performance trade-offs. An important distinction is that the two mitigation strategies do not target the same source of disparity. Fairness-aware regularization reduces disparities by directly constraining model outcomes, regardless of whether these differences originate from protected attributes or from labour-market-relevant characteristics. Counterfactual averaging operates differently. By averaging predictions across alternative intersectional identity assignments, it neutralizes the direct influence of protected characteristics while preserving the predictive contribution of qualifications, employment histories, and other job-relevant information. Consequently, some distributional disparities remain even after counterfactual adjustment. However,

these residual differences are more likely to reflect labour-market-relevant heterogeneity than demographic sensitivity itself. This distinction helps explain why counterfactual averaging achieves meaningful reductions in occupation-level disparities while maintaining comparatively strong predictive performance.
Taken together, these findings suggest that technical mitigation should not be understood as a mechanism for eliminating fairness concerns. Rather, fairness interventions modify how predictive influence, demographic sensitivity, and occupational concentration patterns are distributed throughout the recommendation system.

6.3 Implications for Deployment, Explainability, and Partcipatory Design in PES

The findings have direct implications for the deployment of AI-supported recommendation systems in PES. In such environments, algorithmic outputs are embedded within broader sociotechnical decision processes involving institutional constraints, professional judgement, and heterogeneous user populations. Recommendations are not implemented automatically but interpreted, contextualized, and potentially contested within counselling interactions. This deployment context has important implications for fairness governance. First, aggregate fairness metrics alone are insufficient for operational monitoring. Occupation-level diagnostics, fairness flags, and robustness analyses provide complementary tools for identifying localized fairness risks that may otherwise remain hidden behind aggregate improvements. Second, the results strengthen the case for integrating explainability functions into fairness-sensitive recommendation systems. Explainability is relevant not merely as a transparency add-on but as a mechanism for understanding how recommendation outputs are generated, how occupational predictions change under fairness interventions, and which qualification-related variables remain influential after mitigation. In particular, explanation interfaces could support the interpretation of flagged occupations, uncertainty communication, and the identification of variables acting as potential proxies for intersectional group membership. Third, the results underline the importance of user-centered and participatory system design in PES. Algorithmic fairness in counselling systems affects not only caseworkers but also jobseekers as active participants in the recommendation process. Future development should therefore investigate how fairness diagnostics and explanation functions can be embedded into interfaces co-designed with end users. Such approaches may improve transparency, interpretability, and perceived procedural legitimacy while ensuring that fairness interventions remain aligned with practical counselling needs. More broadly, the results suggest that responsible AI deployment in labour-market institutions requires moving beyond purely technical fairness optimization toward iterative, human-centered, and governance-oriented fairness management. In this sense, intersectional fairness analysis

should not be interpreted as a one-time evaluation exercise but as an ongoing component of deployment, monitoring, and participatory system development.

7 Conclusion

This paper developed a multi-level framework for evaluating intersectional fairness in AI-supported job matching. Using large-scale administrative labour-market data and a multiclass recommendation setting with 141 occupational groups, we distinguished between distributional fairness, predictive fairness, and residual operational fairness risk. Our findings demonstrate that fairness in employment-related AI systems cannot be reduced to a single metric or analytical level. The unconstrained baseline model does not merely reproduce observed labour-market structures, but systematically reshapes occupational outcome distributions across intersectional groups. Fairness-aware regularization improves predictive fairness as measured by the equal opportunity gap, but introduces pronounced trade-offs with predictive performance and does not eliminate distributional distortions. Rather than removing inequality, technical mitigation primarily compresses and redistributes fairness risks across occupations. The occupation-level fairness flag and the counterfactual robustness analysis further show that residual fairness risks persist even under strong regularization. Some occupational fields remain unusually sensitive to intersectional composition shifts or identity perturbations, indicating that aggregate fairness improvements may coexist with localized distributional distortions. These findings underscore the importance of complementing aggregate fairness metrics with diagnostic mechanisms capable of identifying where residual risks remain concentrated.

Limitations of this study should be considered in interpreting these results. The analysis is conditioned by the structure of administrative labour-market data, which – while enabling large-scale intersectional analysis through variables such as age, gender, and nationality – does not systematically capture other socially salient dimensions, including disability status or racialised experiences beyond nationality. Gender is additionally recorded in a binary format, which reflects institutional classification practices and limits the representation of gender diversity within the data. Moreover, the empirical focus on a specific class of fairness-aware regularisation strategies and evaluation metrics provides a structured basis for comparison, but does not exhaust the broader space of possible fairness definitions or intervention approaches.

The implications extend beyond model optimization. In public employment services, algorithmic recommendations are embedded within sociotechnical decision environments involving caseworkers, institutional constraints, and jobseekers as active participants in the

counselling process. Fairness evaluation should therefore not be understood as a purely technical exercise, but as part of a broader governance challenge that includes explainability, monitoring, and human oversight.

Future research should investigate how intersectional fairness diagnostics can be operationalized within deployment-oriented AI systems. This includes the development of explanation interfaces that support both caseworkers and jobseekers, uncertainty communication mechanisms for fairness-sensitive recommendations, and participatory design approaches that involve end users throughout system development, evaluation, and governance. More broadly, our results suggest that responsible AI deployment in labour-market institutions requires moving beyond static fairness metrics toward iterative, human-centered, and context-sensitive fairness governance.

References

Agbasiere, Chinyere; Nze-Igwe, Goodness (2025): Algorithmic Fairness in Recruitment: Designing AI-Powered Hiring Tools to Identify and Reduce Biases in Candidate Selection. In: Path of Science, Vol. 11, No. 4, p. 5001–5021, URL https://pathofscience.org/index.php/ps/article/view/3471.

An, Jiafu; Huang, Difang; Lin, Chen; Tai, Mingzhu (2025): Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. In: PNAS Nexus, Vol. 4, No. 3, p. pgaf089, URL https://doi.org/10.1093/pnasnexus/pgaf089.

Bach, Ruben L.; Kern, Christoph; Mautner, Hannah; Kreuter, Frauke (2023): The impact of modeling decisions in statistical profiling. In: Data & Policy, Vol. 5, p. e32.

Bauer, Bernhard; Mühlbauer, Sabrina; Schlögl-Flierl, Kerstin; Weber, Enzo; Ziethmann, Paula (2025): Ethical Integration in Public Sector AI. The Case of Algorithmic Systems in the Public Employment Service in Germany. In: IAB Discussion Paper, , No. 12.

Bied, Guillaume; Gaillac, Christophe; Hoffmann, Morgane; Caillou, Philippe; Crépon, Bruno; Nathan, Solal; Sebag, Michele (2023): Fairness in job recommendations: estimating, explaining, and reducing gender gaps. In: Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Vol. 3523, Krakow, Poland: CEUR-WS.org, URL https://inria.hal.science/hal-04438512.

Brioscú, Ailbhe; Lauringson, Anne; Saint-Martin, Anne; Xenogiani, Theodora (2024): A new dawn for public employment services: Service delivery in the age of artificial intelligence. In: OECD Artificial Intelligence Papers, Vol. 19.

Broecke, Stijn (2023): Artificial intelligence and labour market matching. In: OECD Publishing, Vol. 284.

Buolamwini, Joy; Gebru, Timnit (2018): Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Friedler, Sorelle A.; Wilson, Christo (Eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Vol. 81 of Proceedings of Machine Learning Research, PMLR, p. 77–91, URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Buyl, Maarten; Cociancig, Christina; Frattone, Cristina; Roekens, Nele (2022): Tackling Algorithmic Disability Discrimination in the Hiring Process: An Ethical, Legal and Technical Analysis. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), New York, NY, USA: Association for Computing Machinery, p. 1071–1082, URL https://doi.org/10.1145/3531146.3533169.

Chen, Zhenpeng; Zhang, Jie M.; Hort, Max; Harman, Mark; Sarro, Federica (2024): Fairness Testing: A Comprehensive Survey and Analysis of Trends. In: ACM Trans. Softw. Eng. Methodol., Vol. 33, No. 5.

Crenshaw, Kimberle Williams (1991): Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. In: Stanford Law Review, Vol. 43, No. 6, p. 1241–1299, URL http://www.jstor.org/stable/1229039.

Davidson, Jason; Schnepp, Jerry; Buttrick, Hilary; Gupta, Ankur (2026): A framework for ethical AI-HRM development and use. In: AI Ethics, Vol. 6, No. 28.

Davis, Kathy (2008): Intersectionality as Buzzword: A Sociology of Science Perspective on What Makes a Feminist Theory Successful. In: Feminist Theory, Vol. 9, No. 1, p. 67–85.

Desiere, Sam; Struyven, Ludo (2021): Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. In: Journal of Social Policy, Vol. 50, No. 2, p. 367–385.

Flügge, Asbjorn Ammitzboll (2021): Perspectives from Practice: Algorithmic Decision-Making in Public Employment Services. In: Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing, CSCW '21 Companion, New York, NY, USA: Association for Computing Machinery, p. 253–255, URL https://doi.org/10.1145/3462204.3481787.

Hancock, Ange-Marie (2007): Intersectionality as a Normative and Empirical Paradigm. In: Politics & Gender, Vol. 3, p. 248–254.

Hill Collins, Patricia; Bildge, Sirma (2020): Intersectionality. Cambridge: Polity.

Körtner, John; Bonoli, Giuliano (2022): Predictive Algorithms in the Delivery of Public Employment Services. In: SocArXiv, p. 19.

Malin, Christine; Fleiß, Jürgen; Ortlieb, Renate; Thalmann, Stefan (2025): Rejected by an AI? Comparing job applicants' fairness perceptions of artificial intelligence and humans in personnel selection. In: Frontiers in Artificial Intelligence, Vol. Volume 8 - 2025, URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1671997.

Mehrotra, Anay; Pradelski, Bary S. R.; Vishnoi, Nisheeth K. (2022): Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, New York, NY, USA: Association for Computing Machinery, p. 599–609, URL https://doi.org/10.1145/3531146.3533124.

Mohammadi, Ehsan; Cai, Yizhou; Novin, Alamir; Vera, Valerie; Soltanmohammadi, Ehsan (2025): Who is a scientist? Gender and racial biases in google vision AI. In: AI Ethics, p. 4993–5010.

Mühlbauer, Sabrina; Weber, Enzo (2026): Machine learning for labor market matching. In: Machine Learning with Applications, p. 100 861.

Mühlbauer, Sabrina; Weber, Enzo (2024): Predicting job match quality: A machine learning approach. Tech. Rep., IAB-Discussion Paper.

Nielsen, Mathias Wullum; Gissi, Elena; Heidari, Shirin; Horton, Richard; Nadeau, Kari C; Ngila, Dorothy; Noble, Safiya Umoja; Paik, Hee Young; Tadesse, Girmaw Abebe; Zeng, Eddy Y; et al. (2025): Intersectional analysis for science and technology. In: Nature, p. 329–337.

Page, Sarah-Jane; Yip, Andrew K.T. (2020): Intersecting Religion and Sexuality: Sociological Perspectives. Leiden, Niederlande: Brill, URL https://brill.com/view/title/38647.

Poe, Robert Lee; El Mestari, Soumia Zohra (2024): The Conflict Between Algorithmic Fairness and Non-Discrimination: An Analysis of Fair Automated Hiring. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), New York, NY, USA: Association for Computing Machinery, p. 1907–1916, URL https://doi.org/10.1145/3630106.3659015.

Sharma, Ravi; Rajput, Akansha; Srivastav, Vaibhav Kumar (2025): Job Description and Resume Matching System Using NLP. In: 2025 International Conference on Next Generation Information System Engineering (NGISE), Vol. 1, p. 1–5.

Suk, Youmi; Han, Kyung (Chris) T (2023): Evaluating Intersectional Fairness in Algorithmic Decision Making Using Intersectional Differential Algorithmic Functioning. In: PsyArXiv.

Vethman, Steven; Smit, Quirine T. S.; van Liebergen, Nina M.; Veenman, Cor J. (2025): Fairness Beyond the Algorithmic Frame: Actionable Recommendations for an Intersectional Approach. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25, New York, NY, USA: Association for Computing Machinery, p. 3276–3290, URL https://doi.org/10.1145/3715275.3732210.

Wang, Ze; Wu, Zekun; Guan, Xin; Thaler, Michael; Koshiyama, Adriano; Lu, Skylar; Beepath, Sachin; Ertekin, Ediz; Perez-Ortiz, Maria (2024): JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models. In: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, p. 3227–3246, URL http://dx.doi.org/10.18653/v1/2024.findings-emnlp.184.

Zollo, Thomas; Rajaneesh, Nikita; Zemel, Richard; Gillis, Talia; Black, Emily (2025): Towards Effective Discrimination Testing for Generative AI. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25, New York, NY, USA: Association for Computing Machinery, p. 1028–1047, URL https://doi.org/10.1145/3715275.3732067.

Appendix

A1 Distribution of Persons over Intersectional Groups

Table A1: Distribution of intersectional groups over persons

Gender, nationality, and age group	Share
Male, Germany, 25–34	12.79
Female, Germany, 25–34	10.50
Male, Germany, 35–44	8.25
Female, Germany, 35–44	7.74
Male, Germany, 15–24	7.74
Male, Germany, 45–54	6.41
Female, Germany, 45–54	6.34
Female, Germany, 15–24	5.90
Male, Germany, 55–64	4.71
Female, Germany, 55–64	4.26
Male, 10 main migration countries, 25–34	3.24
Male, 10 main migration countries, 35–44	2.26
Male, Remaining, 25–34	1.61
Female, 10 main migration countries, 25–34	1.48
Male, 10 main migration countries, 45–54	1.46
Male, 10 main migration countries, 15–24	1.40
Female, 10 main migration countries, 35–44	1.40
Male, Remaining, 35–44	1.12
Female, 10 main migration countries, 45–54	1.00
Female, Remaining, 35–44	0.77
Female, Remaining, 25–34	0.71
Female, 10 main migration countries, 15–24	0.62
Male, EU, 25–34	0.59
Male, Remaining, 45–54	0.59
Male, Remaining, 15–24	0.54
Male, 10 main migration countries, 55–64	0.52

Anmerkung: Text.

Source: ©IAB

A2 JSD Complete for Original Data

Table A2: Jensen-Shannon divergence in the observed occupational data

JSD	Class count	Occupational group
0.6181	112	Aircraft pilots
0.5483	29	Occupations in musical instrument making
0.5395	46	Occupations in fishing
0.5325	10210	Doctors' receptionists and assistants
0.5226	57	Legislators and senior officials of special interest organisations
0.5183	497	Occupations in floristry
0.5087	1044	Floor layers
0.5084	247	Occupations in veterinary medicine and non-medical animal health practitioners
0.4952	2997	Occupations in plumping, sanitation, heating, ventilating, and air conditioning
0.4884	131	Ship's officers and masters
0.4874	9303	Occupations in building construction
0.4870	897	Occupations in psychology and non-medical psychotherapy
0.4852	25	Occupations in philology
0.4840	43	Occupations in vini- and viticulture
0.4777	5871	Occupations in housekeeping and consumer counselling
0.4764	2403	Occupations in civil engineering
0.4668	3228	Occupations in the interior construction and dry walling, insulation, carpentry, glazing, roller shutter and jalousie installation
0.4598	4150	Drivers and operators of construction and transportation vehicles and equipment
0.4590	6538	Occupations in building services engineering
0.4541	5159	Occupations in metal constructing and welding
0.4520	176	Occupations in underground and surface mining and blasting engineering
0.4483	278	Occupations in environmental protection engineering
0.4473	157	Occupations in stage, costume and prop design,
0.4459	170	Occupations in police and criminal investigation, jurisdiction and the penal institution
0.4397	4752	Technical occupations in energy technologies
0.4394	1522	Sales occupations (retail) selling drugstore products, pharmaceuticals, medical supplies and healthcare goods
0.4310	794	Drivers of vehicles in railway traffic

Continued on next page

JSD	Class count	Occupational group
0.4292	35	Artisans designing ceramics and glassware
0.4281	160	Occupations in beverage production
0.4261	2830	Occupations in building services and waste disposal
0.4230	644	Occupations in the inspection and maintenance of traffic infrastructure
0.4226	134	Sales occupations (retail) selling books, art, antiques, musical instruments, recordings or sheet music
0.4208	3341	Occupations in legal services, jurisdiction, and other officers of the court
0.4198	7200	Technical occupations in the automotive, aeronautic, aerospace and ship building industries
0.4188	1000	Conditioning and processing of natural stone and minerals, production of building materials
0.4174	328	Technical occupations in railway, aircraft and ship operation
0.4144	4352	Painters and varnishers, plasterers, occupations in the waterproofing of buildings, preservation of structures and wooden building components
0.4128	25634	Driver of vehicles in road traffic
0.4100	3745	Occupations in body care
0.4044	1663	Occupations in mechatronics, automation and control technology
0.4038	66	Occupations in economics
0.4036	972	Occupations in colour coating and varnishing
0.3988	550	Occupations in forestry, hunting and landscape preservation
0.3977	171	Musicians, singers and conductors
0.3969	227	Occupations in horsekeeping
0.3906	102	Occupations in mathematics and statistics
0.3906	936	Occupations in metal-making
0.3880	953	Occupations in precision mechanics and tool making
0.3878	666	Occupations in public relations
0.3787	187	Occupations in geology, geography and meteorology
0.3784	4259	Occupations in human medicine and dentistry
0.3756	3794	Occupations in wood-working and -processing
0.3727	481	Occupations in treatment of metal surfaces
0.3701	114	Occupations in the humanities
0.3651	396	Occupations providing nutritional advice or health counselling, and occupations in wellness
0.3638	3969	Occupations in non-medical therapy and alternative medicine

Continued on next page

JSD	Class count	Occupational group
0.3627	250	Occupations in funeral services
0.3524	1278	Occupations in editorial work and journalism
0.3487	323	Occupations in environmental protection management and environmental protection consulting
0.3466	23235	Occupations in cleaning services
0.3459	2368	Occupations in technical research and development
0.3459	3575	Occupations in software development and programming
0.3427	1672	Managing directors and executive board members
0.3424	16986	Occupations in machine-building and -operating
0.3419	1981	Occupations in pharmacy
0.3419	797	Occupations in traffic surveillance and control
0.3410	82	Occupations in artisan craftwork and fine arts
0.3331	1139	Teachers for occupation-specific subjects at vocational schools and in-company instructors in vocational training
0.3308	29492	Occupations in education and social work, and pedagogic specialists in social care work
0.3300	816	Occupations in event technology, cinematography, and sound engineering
0.3278	219	Occupations in product and industrial design
0.3277	2676	Occupations in IT-network engineering, IT-coordination, IT-administration and IT-organisation
0.3248	209	Occupations in industrial ceramic-making and -processing
0.3238	12155	Occupations in public administration
0.3167	772	Occupations in biology
0.3160	3705	Teachers in schools of general education
0.3126	1122	Laboratory occupations in medicine
0.3121	12110	Occupations in metalworking
0.3120	431	Occupations in physics
0.3098	625	Occupations in media, documentation and information services
0.3086	4526	Occupations in human resources management and personnel service
0.3070	400	Occupations in animal care
0.3062	7589	Sales occupations (retail) selling foodstuffs
0.3057	380	Occupations in interior design, visual marketing, and interior decoration
0.3036	856	Occupations in the production of clothing and other textile products
0.3033	12145	Occupations in geriatric care

Continued on next page

JSD	Class count	Occupational group
0.3030	570	Occupations in industrial glass-making and -processing
0.3028	81	Military
0.3022	34068	Office clerks and secretaries
0.3021	2989	Teachers and researcher at universities and colleges
0.3009	156	Presenters and entertainers
0.3002	415	Occupations in publishing and media management
0.2946	421	Occupations in theatre, film and television productions
0.2922	3725	Occupations in computer science
0.2896	1419	Teachers at educational institutions other than schools (except driving, flying and sports instructors)
0.2892	169	Technical and management occupations in museums and exhibitions
0.2835	12956	Occupations in nursing, emergency medical services and obstetrics
0.2833	3124	Occupations in IT-system-analysis, IT-application-consulting and IT-sales
0.2812	206	Occupations in theology and church community work
0.2783	242	Occupations in occupational health and safety administration, public health authority, and disinfection
0.2771	2117	Occupations in tax consultancy
0.2768	287	Occupations in surveying and cartography
0.2766	4956	Occupations in hotels
0.2755	200	Occupations in photography and photographic technology
0.2746	6552	Occupations in accounting, controlling and auditing
0.2656	6865	Occupations in electrical engineering
0.2653	4455	Occupations in plastic- and rubber-making and -processing
0.2649	5717	Occupations in gardening
0.2648	827	Technical occupations in paper-making and -processing and packaging
0.2582	47541	Occupations in warehousing and logistics, in postal and other delivery services, and in cargo handling
0.2547	2154	Occupations in the social sciences
0.2507	1119	Occupations in tourism and the sports (and fitness) industry
0.2506	1635	Trading occupations
0.2432	172	Artisans working with metal
0.2410	14129	Cooking occupations
0.2410	2637	Draftspersons, technical designers, and model makers
0.2384	2277	Occupations in technical media design

Continued on next page

JSD	Class count	Occupational group
0.2376	5372	Occupations in the production of foodstuffs, confectionery and tobacco products
0.2363	844	Occupations in event organisation and management
0.2343	5221	Technical occupations in production planning and scheduling
0.2295	2948	Occupations in construction scheduling and supervision, and architecture
0.2278	320	Occupations in animal husbandry
0.2277	10587	Occupations in physical security, personal protection, fire protection and workplace safety
0.2260	1904	Occupations in farming
0.2244	10680	Occupations in advertising and marketing
0.2207	4122	Occupations in insurance and financial services
0.2206	631	Actors, dancers, athletes and related occupations
0.2204	29441	Sales occupations in retail trade (without product specialisation)
0.2193	1572	Occupations in real estate and facility management
0.2173	282	Occupations in leather- and fur-making and -processing
0.2168	1723	Driving, flying and sports instructors at educational institutions other than schools
0.2151	557	Occupations in textile making
0.2098	12535	Occupations in purchasing and sales
0.2051	790	Occupations in printing technology, print finishing, and book binding
0.2050	19909	Occupations in business organisation and strategy
0.2022	1738	Technical occupations in medicine, orthopaedic and rehabilitation
0.1926	15976	Gastronomy occupations
0.1841	7160	Sales occupations (retail trade) selling clothing, electronic devices, furniture, motor vehicles and other durables
0.1713	3285	Management assistants in transport and logistics
0.1663	768	Service occupations in passenger traffic
0.1517	3573	Occupations in chemistry

Note: The table reports the Jensen-Shannon divergence between the intersectional group distribution within each occupation and the overall population distribution. Higher values indicate stronger structural concentration in the observed data. Occupations are sorted in descending order of JSD.

Source: own calculations. ©IAB

List of Figures

List of Tables

Imprint

IAB-Discussion Paper 06|2026

Publication Date
September 16, 2026

Publisher
Institute for Employment Research
of the Federal Employment Agency
Regensburger Straße 104
90478 Nürnberg
Germany

Download
https://doku.iab.de/discussionpapers/2026/dp0626.pdf

All publications in the series "IAB-Discussion Paper" can be downloaded from
https://iab.de/en/publications/iab-publications/iab-discussion-paper-en/

Website
https://iab.de/en

ISSN
2195-2663

DOI
10.48720/IAB.DP.2606

Corresponding author
Sabrina Mühlbauer
Telefon 0911 1799743
E-Mail sabrina.muehlbauer@iab.de