From the Fair Distribution of Predictions to the Fair Distribution of Social Goods: Evaluating the Impact of Fair Machine Learning on Long-Term Unemployment

ANONYMOUS AUTHOR(S)

Algorithmic fairness focuses on the distribution of *predictions* at the time of *training*, rather than the distribution of *social goods* that arises after *deploying* the algorithm in a concrete social context. However, requiring a 'fair' distribution of predictions may undermine efforts at establishing a fair distribution of social goods. Our first contribution is conceptual: we argue that addressing the fundamental question that motivates algorithmic fairness requires a notion of *prospective* fairness that anticipates the change in the distribution of social goods after deployment. Our second contribution is theoretical: we provide conditions under which this change is identified from pre-deployment data. That requires distinguishing between, and accounting for, different kinds of performative effects. In particular, we focus on the way predictions change policy decisions and, therefore, the distribution of social goods. Throughout, we are guided by an application from public administration: the use of algorithms to (1) predict who among the recently unemployed will remain unemployed in the long term and (2) target them with labor market programs. Our final contribution is empirical: using administrative data from the Swiss public employment service, we simulate how such policies would affect gender inequalities in long-term unemployment. When risk predictions are required to be 'fair', targeting decisions are less effective, undermining efforts to lower overall levels of long-term unemployment and to close the gender gap in long-term unemployment.

CCS Concepts: • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics → Computing / technology policy**.

Additional Key Words and Phrases: Algorithmic Fairness, Inequality, Prospective Fairness, Active Labor Market Programs

ACM Reference Format:
Anonymous Author(s). 2024. From the Fair Distribution of Predictions to the Fair Distribution of Social Goods: Evaluating the Impact of Fair Machine Learning on Long-Term Unemployment. In . ACM, New York, NY, USA, 28 pages. https://doi.org/XXXXXXX.XXXXXXX

1 A FUNDAMENTAL QUESTION FOR FAIR MACHINE LEARNING

Research in algorithmic fairness is often motivated by the worry that machine learning algorithms will reproduce or exacerbate the structural inequalities reflected in their training data [57, 71]. Indeed, whether an algorithm exacerbates an existing social inequality is emerging as a central compliance criterion in EU non-discrimination law [74]. However, the methodological solutions developed by researchers in algorithmic fairness are, surprisingly, ill-suited for addressing this fundamental question. At some level, the questions of algorithmic fairness are ill-posed: often, it does not make sense to talk about the fairness of a predictor, independent of the policy context in which it is deployed. It is our policies and their effects that are just or unjust; 'fair' predictors can both support unjust policies and undermine just policy. For example, public employment services use predictions of the risk of long-term unemployment to decide who is given access to training programs. Policy doves target those at the highest risk with training programs, while hawks, considering those at the highest risk to be hopeless cases, withhold training on grounds of 'efficiency'. It is clear that the social consequences of errors in prediction differ significantly depending on how these predictions will be used. It would be surprising if we could say whether a predictor is fair independent of this policy context. Therefore, rather than focusing on the distribution of *predictions* at the time of *training*, we focus on the distribution of *social goods* induced by *deploying* a predictive algorithm in a policy context.

2024. Manuscript submitted to ACM

The field of algorithmic fairness has produced many mathematical demonstrations of necessary trade-offs between different notions of 'fairness', and between 'fair' and accurate prediction [14, 20, 43, 59]. This lends the field an air of tragedy and makes the pursuit of fairness seem fundamentally quixotic. But, while mathematical trade-offs exist between predictive accuracy and the 'fair' distribution of predictions, predictive accuracy does not necessarily trade-off against the fair distribution of social goods [23, 69]. Indeed, we should expect that accurate predictions help us to effectively implement policy aimed at ameliorating unjust inequalities. In our empirical case study, we demonstrate that (1) requiring risk predictions to be fair undermines efforts to lower overall levels of long-term unemployment and to close the gender gap in long-term unemployment, (2) that the hawkish policy of withholding training programs from those at the highest risk is no more efficient than the dovish policy of prioritizing those with the highest risk, and (3) that accurate prediction of *counterfactual* treatment outcomes, rather than risk scores, enables individualized targeting and therefore, a better and more equitable distribution of social goods.

Of course, this shift in focus poses methodological challenges. To anticipate the causal effects of embedding a predictive algorithm into a social process, we must make some effort to, first, identify the contextually relevant inequalities in the distribution of social goods; second, understand the policy processes and decisions that partially give rise to, and could conceivably ameliorate, these inequalities; and third, model how algorithmic predictions might *change* these processes and, therefore, the distribution of social goods. Standard algorithmic fairness methods neglect every part of this process [34, 68]. All of these methods impose constraints on predictions that hold in the (retrospective) training distribution. By focusing on the distribution of predictions at the time of training, they obscure substantive inequalities in real-world quantities and neglect the changes in decision-making that arise from the deployment of predictive algorithms. Consequently, these methods fail to anticipate the effects of *deploying* these algorithms on the distribution of social goods. We address these shortcomings in the following way:

- We reconceptualize algorithmic fairness questions as policy problems. *Prospective fairness* is a matter of anticipating the effect of deploying an algorithmically informed policy on inequality in social goods.
- We state formal conditions under which the effect of deploying an algorithmically informed policy on context-relevant inequalities is identified from pre-deployment data.
- We illustrate our approach with a case study on the statistical profiling of registered unemployed using a rich administrative dataset from Switzerland. We study the likely effects of two algorithmic policy proposals on the gender gap in the rates of long-term unemployment.

Our case study is based on administrative data from the Swiss Active Labor Market Policy Evaluation Dataset. The original sample, collected in 2003, contains roughly one hundred thousand observations of registered unemployed aged 24 to 55. Although most unemployed were not assigned to any program, we observe outcomes for six labor market programs. The Swiss labor market, as outlined in Section 3, is characterized by an overall unemployment rate of about 4%, a high rate of long-term unemployment (LTU), and a persistent gender reemployment gap (2). In the administrative data, the LTU gender gap is at 3.9%, with an LTU rate of 43.6% among women and 39.7% among men. The gap between Swiss citizens and non-citizens is at 15.8%, with a rate of 35.7% among Swiss citizens and 51.5% among non-citizens.

The plan of the paper is as follows: first, we argue for *prospective fairness* as a conceptual framework and survey related work; section 3 introduces two recently proposed algorithmic policies intended to support public employment agencies in reducing long-term unemployment; we argue that, in this context, the gender gap in long-term unemployment is a simple and intuitive measure of systemic inequality; section 4 formalizes conditions under which the causal effect of deploying an algorithmically informed policy on a measure of systematic inequality is identified from pre-deployment data; in section 5 we illustrate the method with an extended case study, simulating two proposed profiling policies and their effects on the gender reemployment gap. Section 6 concludes and outlines directions for future work.

2 FROM RETROSPECTIVE TO PROSPECTIVE FAIRNESS

In paradigmatic risk-assessment applications, machine learners are concerned with learning a function that takes as input some features X and a sensitive attribute A and outputs a score R which is valuable for predicting an outcome Y. The algorithmic score R is meant to inform some important decision D that, typically, is causally relevant for the outcome Y. In the application that concerns us in this paper, features such as the education and employment history (X) and gender (A) of a recently unemployed person are used to compute a risk score (R) of long-term unemployment (Y). This risk score R is meant to support a caseworker at a public employment agency in making a plan (D) about how to re-enter employment. This plan may be as simple as requiring the client to apply to some minimum number of jobs every month or referring them to one of a variety of job-training programs.

Formal fairness proposals require that some property is satisfied by either the joint distribution $P(A, X, R, D, Y)$ or the causal structure G giving rise to it. Individual fairness proposals introduce a similarity metric M on (A, X) and suggest that similar individuals should have similar risk scores. In all these cases, the relevant fairness property is a function $\varphi(P, G, M)$. Group-based fairness [8] ignores all but the first parameter; causal fairness [41, 50] ignores the last; and individual fairness [30] ignores the second. All these proposals agree that fairness is a function of the distribution (and perhaps the causal structure) at the time when the prediction algorithm has been trained, *but before it has been deployed*. We claim that addressing the fundamental question of fair machine learning requires comparing the status quo *before* deployment with the situation likely to arise *after* deployment. In other words: *prospective* fairness is a matter of anticipating the change from $\varphi(P_{\text{pre}}, D_{\text{pre}}, M)$ to $\varphi(P_{\text{post}}, D_{\text{post}}, M)$. We do not claim that there is a single correct inequality measure $\varphi(\cdot)$, nor even that there is an all-things-considered way of trading off different candidates, only that we must make a good faith effort to anticipate changes in the relevant measures of inequality.

As shown in Figure 1, deploying a decision support algorithm introduces a causal path from the predicted risk score R to the decision D. Importantly, the outcome variable Y is causally downstream of this intervention. The addition of a causal path is modeled as a *structural* intervention [17, 58].

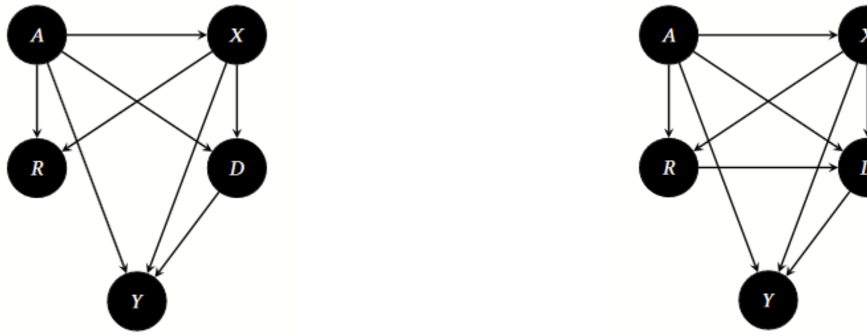
From a dynamical perspective, static and retrospective fairness proposals go wrong in two ways. In the worst case, they are *self-undermining*: satisfying the fairness criteria at the time of training necessitates violating them after implementation. For example, Mishler and Dalmasso [60] show that satisfying the fairness notions of sufficiency ($Y \perp A \mid R$) or separation ($R \perp A \mid Y$) at the time of training virtually ensures that they will be violated after deployment. Illustrating the point in terms of sufficiency, where $\perp$ denotes (conditional) statistical independence:

$$Y \perp_{\text{pre}} A \mid R \quad \text{entails} \quad Y \not\perp_{\text{post}} A \mid R.$$

Group-based notions of fairness like sufficiency and separation fall victim to *performativity*: the tendency of an algorithmic policy intervention to shift the distribution away from the one on which it was trained [64]. But as Mishler and Dalmasso [60] show, they are undermined not by an unintended and unforeseen performative effect, but by the *intended, and foreseen* shift in distribution induced by algorithmic support, i.e.:

$$P_{\text{pre}}(D \mid A, X, R) \neq P_{\text{post}}(D \mid A, X, R).$$

In other words, they are undermined by the fact that algorithmic support changes decision-making, which, presumably, is the point of algorithmic support in the first place. Since the distribution of the outcome Y will change after deployment,



(a) Causal structure G_{pre} *before* deploying an algorithmically informed policy.

(b) Causal structure G_{post} *after* deploying an algorithmically informed policy.

Fig. 1. The left hand side shows the pre-deployment causal graph G_{pre} inducing a joint probability distribution P_{pre} over sensitive attributes A, features X, risk score R, decision D, and outcome variable Y. The risk score R is the output of a learned function from A and X. Since this graph represents the situation after training, but before deployment, there is no arrow from the risk score R to the decision D. *Retrospective* fairness formulates constraints $\varphi(G_{\text{pre}}, P_{\text{pre}}, M)$ on the pre-deployment arrangement alone. The right-hand side represents the situation after the algorithmically informed policy has been deployed, with predictions R now affecting decisions D. Prospective fairness requires comparing the consequences of intervening on the structure of G_{pre} and moving to G_{post}. In other words, comparing $\varphi(G_{\text{pre}}, P_{\text{pre}}, M)$ with $\varphi(G_{\text{post}}, P_{\text{post}}, M)$.

Berk et al. [11] advises against group-based metrics involving it, opting for statistical parity ($R \perp A$) instead. Of course, independence requires a loss of predictive accuracy, which may undermine even the most benevolent policies.

It is not likely that individual and causal fairness proposals are so drastically self-undermining. So long as the similarity metric stays constant, an algorithm that treats similar people similarly will continue to do so after deployment. If, as Kilbertus et al. [41] suggest, causal fairness is a matter of making sure that all paths from the sensitive attribute A to the prediction R are appropriately mediated, then causal fairness is safe from performative effects so long as the qualitative causal structure *upstream* of the prediction R remains constant.

But even if causal and individual fairness proposals are not so dramatically self-undermining, they are simply *not testing* whether the algorithm reproduces or exacerbates inequalities in social goods, since the distribution of social goods is causally *downstream* of algorithmic predictions. In particular, it is customary to ignore the real-world dependence between A and Y induced by the social status quo as the target of an intervention, since nothing can be done about it at the time of training. Instead, fairness researchers focused on whether the risk score *itself* is fair, whether in the group, individual, or causal sense. However, from the dynamical perspective, it is perfectly reasonable to ask whether the proposed algorithmic policy will exacerbate the systemic inequality reflected in the dependence between gender (A) and long-term unemployment (Y). Indeed, simple dynamical models and simulations suggest that algorithms meeting static fairness notions at training may exacerbate inequalities in outcomes in the long run [56, 75]. Streamlined dynamical models and simulations are a valuable tool in evaluating the long-run effects of fairness-constrained algorithms. The dual contributions of this paper are (1) a theoretical result giving conditions under which the effect of deploying the algorithmic policy on the joint distribution of (Y, A) is identified and (2) a realistic case study that forecasts, from administrative data, the effects of algorithmic policies in public employment on the joint distribution of gender and long-term unemployment.

2.1 Related Work

In machine learning, the fairness debate began with risk assessment tools for decision- and policy-making [5, 20, 43, 61]. To this day, many standard case studies e.g., lending, school admissions, and pretrial detention, fall within this scope. See Berk et al. [10] for a review on fairness in risk assessment and Borsboom et al. [14] and Hutchinson and Mitchell [37] for predecessors in psychometrics. Since then, researchers have stressed the importance of explicitly differentiating policy decisions from the risk predictions that inform them [7, 9, 49, 70] and of studying machine learning algorithms in their socio-technological contexts [68]. We incorporate both of these insights into the present work.

A central negative result emerging from recent fairness literature highlights the dynamically self-undermining nature of group-based fairness constraints that include the outcome variable Y. Mishler and Dalmasso [60] show that a classifier that is formally fair in the training distribution will violate the respective fairness constraint in the post-deployment distribution. Coston et al. [24] suggests that the group-based fairness notion be formulated instead in terms of the potential outcomes Y^d. These alternative proposals are no longer self-undermining, but they are still not testing the policy's effect on inequality in the distribution of social goods. This paper builds upon the negative results of Berk et al. [11] and Mishler and Dalmasso [60]: we show how the post-interventional effect of an algorithmically informed policy on the distribution of social goods can be identified from a combination of (1) observational, pre-deployment data and (2) models of the policy proposal.

An emerging literature on long-term fairness focuses on the dynamic evolution of systems under sequential-decision making, static fairness constraints, and feedback loops; see Zhang and Liu [75] for a survey. Ensign et al. [31] consider predictive feedback loops from selective data collection in predictive policing. Hu and Chen [36] propose short-term interventions in the labor market to achieve long-term objectives. Using two-stage models, Liu et al. [56] and Kannan et al. [38] show that retrospective fairness constraints can, under some conditions, have negative effects on outcomes in disadvantaged groups. With simulation studies, D'Amour et al. [27] and Zhang et al. [76] confirm that imposing static fairness constraints does not guarantee that these constraints are met over time and can, under some conditions, exacerbate inequalities in social goods. Scher et al. [66] model long-term effects of statistical profiling for the allocation of unemployed into labor market programs on skill levels. The picture emerging from this literature is that post-interventional outcomes of algorithmic policies are a relevant dimension for normative analysis that is not adequately captured by retrospective fairness notions designed to hold in the training distribution.

3 STATISTICAL PROFILING OF THE UNEMPLOYED

Since the 1990s, participation in active labor market programs (ALMPs) has been a condition for receiving unemployment benefits in many OECD countries [22]. ALMPs take many forms, but paradigmatic examples include resume workshops, job-training programs, and placement services, see Bonoli [13] for a helpful taxonomy. Evaluations of ALMPs across OECD countries find small but positive effects on labor market outcomes [18, 52, 73]. Importantly, the literature also reports large effect-size heterogeneity between programs and demographics, as well as assignment strategies that are as good as random for Switzerland [46], Belgium [21], and Germany [33]. This implies potential welfare gains from a more targeted allocation into programs, especially when taking into account opportunity costs—a compelling motivation for algorithmic support. Indeed, the subsequent case study suggests that, if allocation decisions are made based on data-driven estimates of individualized treatment effects, the gender reemployment gap, as well as overall long-term unemployment, can be significantly reduced.

Statistical profiling of the unemployed is current practice in various OECD countries including Australia, the Netherlands, and Flanders, Belgium [28]. Paradigmatically, supervised learning techniques are employed to predict who is at risk of becoming long-term unemployed (LTU) [62]. Such tools are regularly framed as introducing objectivity and effectiveness in the provision of public goods and align with demands for evidence-based policy and digitization in public administration. ALMPs target *supply-side* problems by increasing human capital and *matching* problems by supporting job search. *Demand-side* policies that focus on the creation of jobs are not considered [34].

Individual scores predicting the risk of long-term unemployment support a variety of decisions. For example, the public employment service (PES) of Flanders so far uses risk scores only to help caseworkers and line managers decide who to contact first, prioritizing those at higher risk [29]. In contrast, the PES of Austria (plans to) use risk scores to classify the recent unemployed into three groups: those with good prospects in the next six months; those with bad prospects in the next two years; and everyone else. The proposed policy of the Austrian PES is to focus support measures on the third group while offering only limited support to the other two. Advocates claim that, since ALMPs are expensive and would not significantly improve the re-employment probabilities of individuals with very good or very bad prospects, considerations of cost-effectiveness require a focus on those with middling prospects [3]. However intuitive this may seem, it is nowhere substantively argued that statistical predictions of long-term unemployment from observational data are reliable estimates for the effectiveness of administrative interventions. One worry is that the unemployed who are labeled high-risk tend to be similar to those who, historically, received ineffective programs. This is further complicated by the presence of long-standing structural inequalities in the labor market, which may be reproduced by algorithmic policies leaving those with “poor prospects" to their own devices. In the subsequent simulation study, the efficiency claims made in favor of Austrian-style policy are not corroborated.

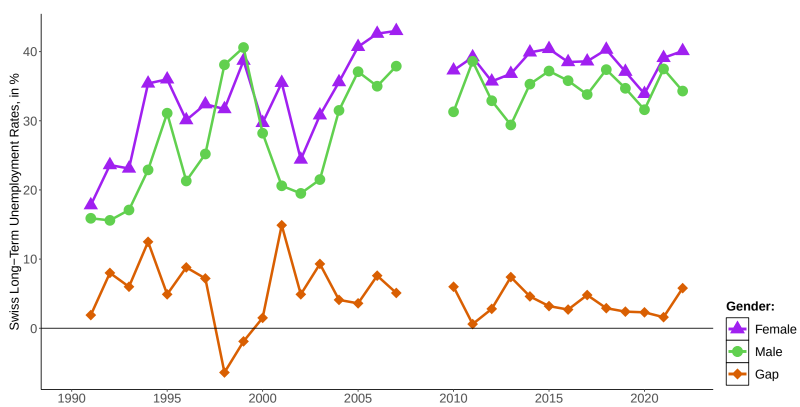


Fig. 2. Swiss Long-Term Unemployment Rates by Gender. Data for the period 2010-2022 are from Eurostat [32], where the gender share of long-term unemployment is computed as the share of all unemployed men/women ages 20-64 who are unemployed for more than a year. Data for the period 1991-2007 are from the 2012 Swiss Social Report [15], where age information is not available. Data for 2008-9 is not readily available.

Labor markets in OECD countries are structured by various inequalities. Gender is a particularly long-standing and significant axis of inequality in labor markets, with the gender pay gap and the child penalty being notorious examples [12, 44]. On the other hand, the gender gap in unemployment rates has largely disappeared over the last decades [2]. Nevertheless, structural differences in unemployment dynamics remain. For example, although women in Germany are less likely to enter into unemployment, their exit probabilities are also lower [16]. Similarly, there is a longstanding gender gap in long-term unemployment in Switzerland (see Figure 2). The obvious worry is that prediction algorithms will pick up on these historical trends, as demonstrated in Kern et al. [40]. The Austrian proposal for an LTU prediction algorithm furnishes a particularly dramatic example. That algorithm takes as input an explicitly gendered feature "obligation to care", which has a negative effect on the predicted re-employment probability and, by design, is only active for women [3]. This controversial design choice was justified as reflecting the "harsh reality" of the gendered distribution of care responsibilities. Whatever the wisdom of this particular variable definition, many other algorithms would pick up on the same historical patterns. Moreover, if the intended use of these predictions is to withhold support for individuals at high risk of long-term unemployment, it is clear that such a policy might exacerbate the situation by further punishing women for greater care obligations. Hopefully, the preceding motivates the need for a prospective fairness methodology that assesses whether women's re-employment probability suffers under a proposed algorithmic policy. More abstractly, what is needed is a way to predict how the pre-deployment probability $P_{\text{pre}}(Y \mid A)$ will compare with the post-deployment probability $P_{\text{post}}(Y \mid A)$. With these estimates in hand, it would also be possible to predict whether the gender reemployment gap is exacerbated, or ameliorated, under a proposed algorithmic policy. The gender gap in reemployment probabilities is one particular choice for a fairness notion $\varphi(\cdot)$. Variations on this simple metric could be relevant in many other settings. For example, gender gaps in hiring, or racial disparities in incarceration could be criteria that an algorithmically informed policy should, minimally, not exacerbate [39]. In the following section, we give general conditions under which the post-deployment change in the joint distribution of the outcome (Y) and the sensitive attribute (A) is identified from pre-deployment data.

4 IDENTIFIABILITY OF THE POST-DEPLOYMENT DISTRIBUTION OF SOCIAL GOODS

Let A, X, R, D, Y be discrete, *observed* random variables. In our example, A represents gender; X represents baseline covariates observed by the public employment service for the registered unemployed; R is an estimated risk of becoming long-term unemployed; D is an allocation decision made by the public employment service and Y is a binary random variable that is equal to 1 if an individual becomes long-term unemployed. For simplicity, we assume that R is a deterministic function of A and X. We write $\mathcal{A}, \mathcal{X}, \mathcal{R}, \mathcal{D}, \mathcal{Y}$ for the respective ranges of these random variables. For $d \in \mathcal{D}$, let Y^d be the potential outcome under policy d, in other words: Y^d represents what the long-term unemployment status of an individual *would have been* if they had received allocation decision d. Naturally, $Y^1, \ldots, Y^{|\mathcal{D}|}$ are not all observed. Our first assumption is a rather mild one; we require that the observed outcome for individuals allocated to d is precisely Y^d :

$$Y = \sum_{d \in \mathcal{D}} Y^d \mathbb{1}[D = d]. \qquad \text{(Consistency)}$$

Consistency is to be interpreted as holding both before and after the algorithmic policy is implemented.

More substantially, we assume that the potential outcomes and decisions are unconfounded given the observed features (A, X) both before and after the intervention:

$$Y^d \perp_t D \mid A, X. \qquad \text{(Unconfoundedness)}$$

Unconfoundedness is a rather strong assumption that requires that the observed features A, X include all common causes of the decision and outcome. In the case of a fully automated algorithmic policy, unconfoundedness holds by design; but usually, risk assessment tools are employed to support human decisions, not fully automate them [55]. Although it is not fated that all factors relevant to a human decision are available to the data analyst, unconfoundedness is reasonable if rich administrative data sets capture most of the information relevant to allocation decisions. For a case in which this assumption fails, see Petersen et al. [65].

We have argued that, in order to address the fundamental question of fair machine learning, one must predict whether implementing the candidate algorithmically informed policy leads to an improvement, or at least no deterioration, in standards of justice. In the running example, this amounts to comparing features of $P_{\text{pre}}(Y \mid A)$ with $P_{\text{post}}(Y \mid A)$. The first distribution is trivial to estimate, but how to estimate $P_{\text{post}}(Y \mid A)$ from pre-deployment data? Here, the fundamental problem is performativity [64]. Our policy intervention will, in all likelihood, change the process of allocation into labor market programs and, thus, change the distribution of outcomes we are interested in. But not all kinds of performativity are equal. Some performative effects are intended and foreseeable. For example, the *algorithmic* effect is the intended change in decision-making due to algorithmic support:

$$P_{\text{pre}}\left(D = d \mid A = a, X = x\right) \neq P_{\text{post}}\left(D = d \mid A = a, X = x\right). \qquad \text{(Algorithmic Effect)}$$

The first term in this inequality is the propensity score which can be directly estimated from training data. The second term cannot be directly estimated *ex-ante*. Nevertheless, it is possible to make reasonable conjectures about the second term given a concrete proposal for how risk scores should inform decisions. For example, if D is binary, we could model the Austrian proposal as providing support so long as the risk score is neither too high nor low:

$$P_{\text{post}}(D = 1 \mid A = a, X = x) = \mathbb{1}\left[l < R(a, x) < h\right].$$

More complex proposals for how risk scores should influence decisions require more careful modeling. The subsequent empirical case study delivers a more realistic model.

Although we allow for algorithmic effects, these cannot be too strong—the policy cannot create allocation options that did not exist before. That is, the risk assessment tools only change allocation probabilities into *existing* programs. Moreover, we assume that the policy creates no unprecedented allocation-demographic combinations:

$$P_{\text{pre}}(D = d \mid A = a, X = x) > 0 \text{ if } P_{\text{post}}(D = d \mid A = a, X = x) > 0. \qquad \text{(No Unprecedented Decisions)}$$

This would be violated if e.g., no women were allocated to some program before the policy change.

Throughout this paper, we assume that no other forms of performativity occur. In particular, we assume that the conditional average treatment effects (CATEs) of the allocation on the outcome are stable across time:

$$P_{\text{pre}}\left(Y^d \mid A = a, X = x\right) = P_{\text{post}}\left(Y^d \mid A = a, X = x\right). \qquad \text{(Stable CATE)}$$

This amounts to assuming that the effectiveness of the programs (for people with $A = a, X = x$) does not change, so long as all that has changed is the way we *allocate* people to programs. In the case study, we assume that conditional average treatment effects are stable under changes to allocation policies, as well as to the total number of places available in (capacities of) each program. This assumption could be violated if e.g., a program works primarily by making some better off only at the expense of others—if everyone were to receive such a program, it would have no effect [25].

While *algorithmic effects* of deployment are intended and, to some degree, foreseeable types of performativity, *feedback* effects that change the covariates are more complicated to model.[1] Following Mishler and Dalmasso [60] and Coston et al. [24], we assume away the possibility of feedback effects, leaving these for future research:

$$P_{\text{pre}}\left(A = a, X = x\right) = P_{\text{post}}\left(A = a, X = x\right). \qquad \text{(No Feedback)}$$

[1]In the classification of Pagan et al. [63], we focus on what they call "Outcome Feedback Loops". In our terminology, performativity is not exhausted by feedback effects.

No Feedback amounts to assuming that the baseline covariates of the recently employed are identically distributed pre- and post-deployment. Strictly speaking, this is false, since the decisions of caseworkers will affect the covariates of those who re-enter employment and some of them will, eventually, become unemployed again. However, since the pool of employed is much larger than the pool of unemployed, the policies of the employment service have much larger effects on the latter than the former. For this reason, we may hope that feedback effects are not too significant.

No Unprecedented Decisions, Stable CATE and No Feedback might fail dramatically if e.g., the deployment of the policy coincided with a major economic downturn. In a serious downturn, the employment service may have to assist people from previously stable industries (violating No Unprecedented Decisions and No Feedback), or employment prospects might deteriorate for everyone (violating Stable CATE). However, the possibility of such exogenous shocks is not a threat to our methodology. We are interested in the *ceteris paribus* effect of the algorithmic policy on structural inequality, not an all-thing-considered prediction of future economic conditions.

We are now in a position to show that, under the assumptions outlined above, it is possible to predict $P_{\text{post}}(Y = y \mid A = a)$ from pre-interventional data and a supposition about $P_{\text{post}}(D = d \mid A = a, X = x)$. That means that we can also predict changes to the overall reemployment probability $P_{\text{post}}(Y = 0)$ as well as the gender reemployment gap $P_{\text{post}}(Y = 1 \mid A = 1) - P_{\text{post}}(Y = 1 \mid A = 0)$. Each of these are natural and important instances of $\varphi(\cdot)$. The proof is deferred to the supplementary material.

Theorem 4.1. *Suppose that Consistency, Unconfoundedness, No Unprecedented Decisions, Stable CATE and No Feedback hold. Suppose also that $P_{post}(A = a) > 0$. Then, $P_{post}(Y = y \mid A = a)$ is given by*

$$\sum_{(x,d)\in\Pi_{post}} P_{pre}(Y = y \mid A = a, X = x, D = d) P_{pre}(X = x \mid A = a) P_{post}(D = d \mid A = a, X = x),$$

where $\Pi_t = \{(x, d) \in \mathcal{X} \times \mathcal{D} : P_t(X = x, D = d \mid A = a) > 0\}$.

Note that the first two terms in the product are identified from pre-deployment data. Given a sufficiently precise proposal for how risk scores influence decisions, it is also possible to model Π_{post} and the last term before deployment. This allows us to systematically compare different (fairness-constrained) algorithms and decision procedures, and arrive at a reasonable prediction of their combined effect on reemployment probabilities (and the gender reemployment gap) before they are deployed. In the following, we show how this approach works in a realistic case study.

5 EMPIRICAL STUDY: LONG-TERM UNEMPLOYMENT IN SWITZERLAND

We are interested in forecasting the effect of using (fair) risk scores to inform program allocation decisions on both the overall risk of long-term unemployment and the gender reemployment gap. We present an extensive case study built on Swiss administrative data to study three questions: do fairness-constrained risk scores improve outcomes? are restrictive, Austrian-style allocation policies more efficient than Flemish-style policies that prioritize people at high risk? and can we improve outcomes with individualized estimates of program effectiveness?

5.1 Methodology

Our analysis proceeds in the following stages: (1) Using double-robust machine learning, we first estimate the effectiveness of each of the programs for all individuals in our test sample. (2) We estimate risk scores for the individuals in our test sample, using fairness-constrained and fairness-unconstrained methods. We implement two fairness constraints: statistical parity and equal opportunity. (3) For each of the risk scores from stage two, we prioritize the individuals in the test sample. The Flanders-style policy prioritizes those at the highest risk. The Austrian prioritization does the same, but only for those in the 70 – 30th risk percentiles; the rest go to the end of the line. (4) For each priority list from stage three, we assign unemployed to programs until program capacity is reached. We model two assignment schemes. The first assigns individuals to programs randomly. The second uses the results of stage one to assign individuals to the program with the highest estimated effectiveness. Additionally, we consider the effect of increasing program capacities. Finally, we summarize the effects of different combinations of choices from steps (2-4) on overall rates of long-term unemployment and the gender-reemployment gap.

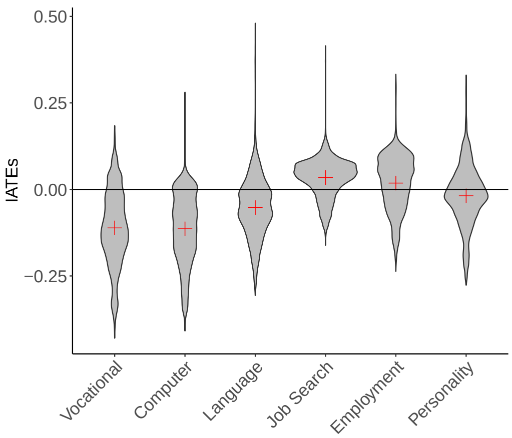
5.1.1 Data. We exploit the administrative Swiss Active Labor Market Policy (ALMP) Evaluation Dataset.[2] The original sample contains observations on 100, 120 registered unemployed in 2003, aged 24 to 55. Recently unemployed received one of seven treatments: *no program, vocational training, computer programs, language courses, job search programs, employment programs, and personality training*. Among the seven treatment options, *no program* and *job search programs* are by far the most common treatments. We restrict the analysis to the German-speaking cantons as assignment strategies differ among the three language regions [45]. To avoid overstating the effectiveness of "no program", we estimate pseudo program starting points for individuals in this treatment arm and exclude those who are re-employed before the pseudo starting point [45, 53]. The final data set contains 64, 296 individuals, which we divide equally into training and test sets. The simulation study is performed on the test set of 32, 148 individuals and all results are reported for this population. Descriptive statistics for the simulation data are reported in Appendix B.1.

For all individuals, we observe employment status for 36 months after registration with the Swiss Public Employment Service (PES). Our target, long-term unemployment, is defined as a binary variable indicating continuous unemployment for 12 months after the (pseudo) program start.[3] The treatment variable is defined as the first program assigned within six months after registering as unemployed. The administrative data includes information on the individual employment biographies, demographics, local labor market conditions as well as information on the individual caseworker and their assessment of their clients' labor market outlook.

5.1.2 Individualized Average Potential Outcomes. We adopt double-robust machine learning for the estimation of individual average potential outcomes (IAPOs) and treatment effects (IATEs) for the seven treatment options [1, 19, 26]. We follow Knaus [45] and Körtner and Bach [51] in their identification strategy and use the R-package causalDML [45]. Inverse probability weighting is used to account for non-random selection into the programs under the identifying assumptions of *Unconfoundedness* (similar to our Unconfoundedness), *Common Support* (No Unprecedented Decisions), and *Stable Unit Treatment Value* (Consistency and Stable CATE). Especially important for the plausibility of Unconfoundedness is the availability of information about the individual caseworker. See Appendix B.3 for a more detailed discussion of the estimation approach.

The resulting (individualized) average treatment effects are given in Figure 3 and Table 3b. They are in line with the results reported in Knaus [45] and Körtner and Bonoli [47]. Vocational Training, Computer Programs, and Language Courses have the strongest effects on reducing (long-term) unemployment. We find that Job Search and Employment Programs on average increase the risk of long-term unemployment by between 2 to 3 percentage points and confirm the high effect heterogeneity in all treatments. The reported treatment effects are the difference of the respective potential outcome scores, where "no program" is the baseline program. IATEs broken down by gender are given in Appendix B.3.

[2]The data is available for scientific use at SWISSbase [54].
[3]This is a deviation from Körtner and Bach [51], who define their target variable as 12 months after registration with the PES.



(a) Individualized Average Treatment Effects.

	ATE	SE	95%-CI
Vocational	-11.12	0.06	[-11.12, -11.12]
Computer	-11.37	0.05	[-11.37, -11.37]
Language	-5.25	0.04	[-5.26, -5.25]
Job Search	3.43	0.03	[3.43, 3.43]
Employment	1.83	0.04	[1.83, 1.83]
Personality	-1.84	0.04	[-1.84, -1.84]

(b) Average Treatment Effects in percentage points, standard errors, and 95% confidence intervals. Negative treatment effects imply a lower risk of becoming long-term unemployed.

Fig. 3. Estimated (Individualized) Average Treatment Effects for the six labor market programs. No program serves as the baseline.

5.1.3 Risk scores. In 2003, program assignment in the Swiss public employment service was made at the discretion of the individual caseworker. This practice continues to this day.[4] For estimating the risk scores to determine a prioritization, all caseworker information is excluded and only data that is reasonably available at registration time is used: characteristics of the unemployed person and the local labor market situation. The sensitive attribute is included as a feature. The full list of features is given in Appendix B.4.

To evaluate the impact of retrospective fairness on the distribution of social goods, we estimate a fairness-unconstrained risk score and two risk scores constrained to satisfy statistical parity[5] and equality of opportunity[6], respectively. Throughout, we use logistic ridge regressions and the R-package FAIRML for estimation of fairness constrained risk scores [67]. We do not require the fairness constraint to be met perfectly.

All three methods, when applying a decision threshold of .5, achieve an accuracy of about 64 − 65%. These results are in line with internationally reported accuracy rates for the prediction of long-term unemployment [28]. The unconstrained risk scores violate statistical parity, with more women than men being predicted to become long-term unemployed (a discrepancy of 0.116). Further, the true (a discrepancy of 0.174) and false positive (0.062) rates are higher for women than for men. The fairness constrained scores reduce these discrepancies. Details on the implementation together with descriptive statistics for the risk scores can be found in Appendix B.4.

5.1.4 Prioritization. For each of the three risk scores from the previous stage, we compile two priority lists modeling the Belgian and Austrian proposals. The Belgian list goes in order of decreasing risk [29]. The Austrian list does the same for those in the $30 - 70th$ risk percentiles. The others are put at the end of the list, in random order [3]. This yields six priority lists, one for each combination of risk score and prioritization scheme.

5.1.5 Program Assignments. For each of the six lists from the previous stage, we assign individuals to programs in order of priority. Individuals are assigned according to two schemes: optimal and random. The first assigns each person

[4]The canton of Freiburg had a pilot study from 2012-2014, providing caseworkers with estimates of the expected length of the unemployment spell [6].
[5]Also called demographic parity or Independence of the predictions from the sensitive attribute [8].
[6]The equality in true positive rates for both groups. This is a relaxation of equalized odds, also called Separation [8].

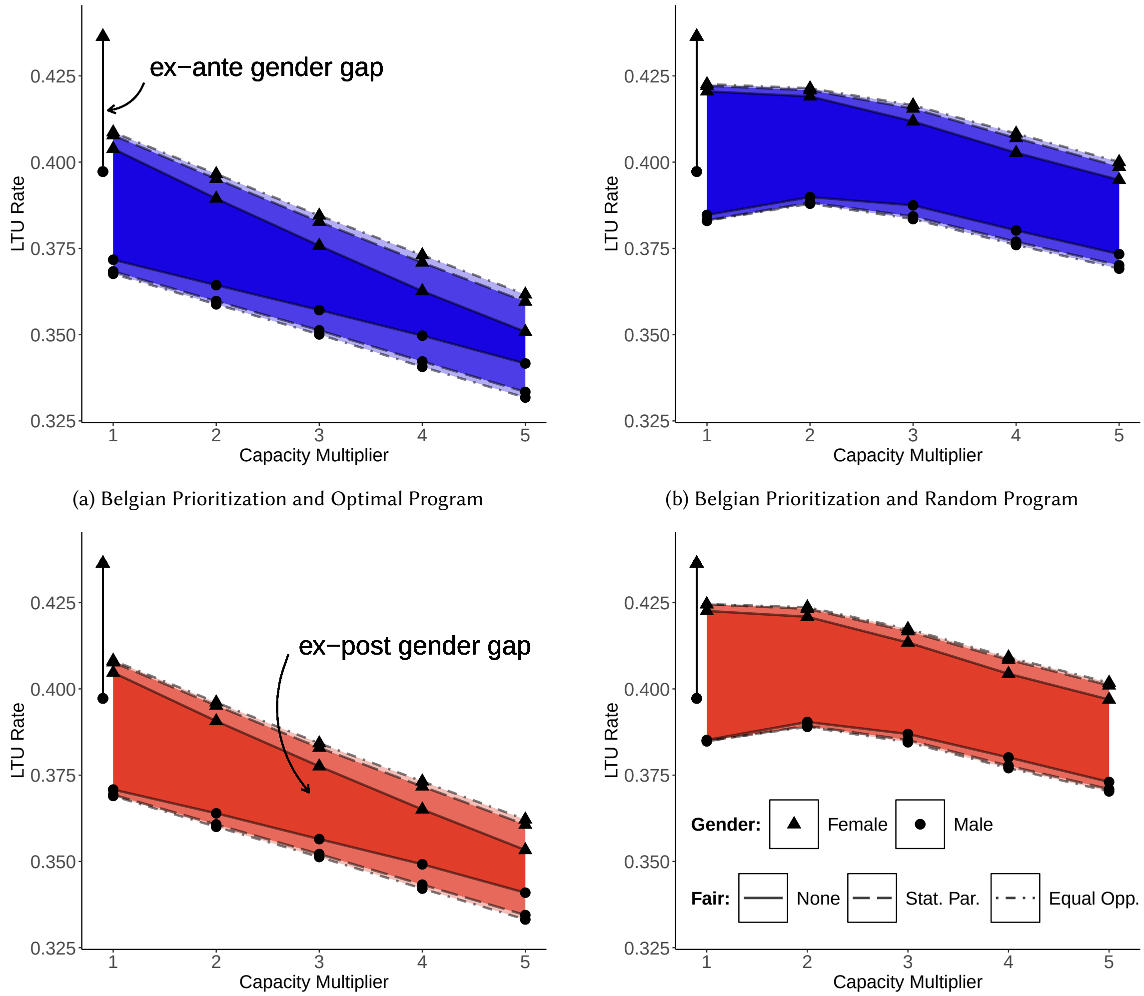
to the program that is most effective for them and not yet at capacity. This models the best-case scenario in which caseworkers are very good at discerning which program is best for each client. The second makes assignments by a uniform draw from the available programs.[7] These two assignment schemes provide upper and lower bounds for what might happen when caseworkers are *informed* by risk scores when making assignment decisions instead of fully automating the decision. To model adjustments to the budget constraint of the PES, we consider the effect of increasing program capacities. As a baseline, we take the program sizes observed in the test set (see Table 1). Then, we consider capacities that are $2-5x$ larger. Because the most effective programs are also the smallest, increasing overall capacities mainly influences outcomes by increasing the capacities of these small but effective programs.

5.2 Results

5.2.1 Fair Prediction and the Fair Distribution of Social Goods. Regardless of the notion of retrospective fairness and the choices made at other stages, constraining risk predictions to be fair yields larger gender reemployment gaps (Figure 4). This is because fairness constraints, by shifting the distribution of risk scores among women to look more like the distribution among men (Figure B.4), tend to underestimate their risk of long-term unemployment. The effect of fairness constraints is to reserve a roughly equal number of seats in effective training programs for men and women. Therefore, fairness-constrained policies induce similar improvements in labor market outcomes for both genders, which keeps the gender reemployment gap relatively constant. On the other hand, fairness unconstrained risk scores are, on average, higher for women. That means that more seats are reserved for women in effective programs–the result is more aggressive reductions in rates of long-term unemployment among women than among men. These effects are only made more pronounced when budget constraints are relaxed and program capacities are increased. For example, at baseline program sizes the combination of Belgian prioritization and individualized treatment decisions yields a 3.2% gender gap in reemployment probabilities (40.4% vs 37.2%) when risk scores are unconstrained and a 4.1% gender gap (40.9% vs 36.8%) when risk scores are constrained to satisfy equal opportunity. This means that, at baseline program sizes, the equal opportunity constraint slightly *exacerbated* the ex-ante gender gap of 3.9% (43.6% vs. 39.7%). If programs are made five times larger, the fairness unconstrained policy reduces the gender gap to .9% (35.1% vs 34.2%) whereas equal opportunity leaves the gender gap relatively unchanged at 3% (36.2% vs 33.2%). All results are given in Tables 5 for baseline capacities and 6 for five-fold capacities. We observe similar patterns for citizenship gaps, reported in Appendix B.6.

5.2.2 Hawks and Doves. Regardless of other choices, the Belgian policy is at least as efficient as the Austrian policy, both in reducing overall rates of long-term unemployment and reducing the gender reemployment gap (Figure 5). This holds both for the optimal program assignment and the random assignment. For example: at baseline program sizes, when the unemployed receive targeted assignment and risk scores are not fairness constrained, the Belgian policy achieves an overall LTU rate of 38.6% and a gender reemployment gap of 3.2% (40.4% vs. 37.2%) whereas the Austrian policy induces an identical overall rate and a gap of 3.4%. If programs are made five times larger, the Belgian policy achieves an overall rate of 34.6% and a gender gap of .9% (35.1% vs 34.2%), whereas the Austrian policy achieves an identical overall rate and a gender gap of 1.2% (35.3% vs 34.1%). Thus, targeting those at the highest risk of long-term unemployment achieves improvements in gender equality without any costs in overall efficiency. A more fine-grained analysis shows that the Belgian prioritization closes the gender gap much more aggressively among married non-citizens, who tend to have the worst labor market outcomes, whereas the Austrian prioritization does slightly better among groups with better

[7] We run this scheme ten times per policy and average over the resulting individual risks for long-term unemployment.



(a) Belgian Prioritization and Optimal Program

(b) Belgian Prioritization and Random Program

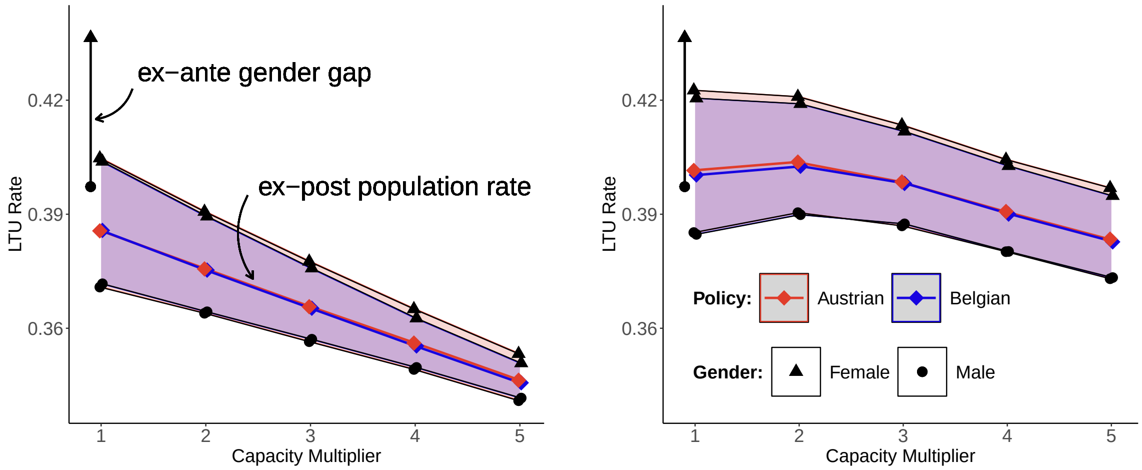
(c) Austrian Prioritization and Optimal Program

(d) Austrian Prioritization and Random Program

Fig. 4. We plot the gender gap in long-term unemployment (LTU) against program capacity for each combination of prioritization and assignment scheme. The level of transparency shows the gender gap for the corresponding fairness constraint: none, statistical parity, or equal opportunity. The unconstrained risk scores (lowest transparency) result in the smallest gender gap. This effect is especially pronounced as program capacity is increased and program assignments are individualized (optimal).

average outcomes B.7. Similar effects are observed for citizenship gaps B.6. Therefore we do not find any efficiency advantage for withholding training from individuals at the highest risk of unemployment.

5.2.3 Gains from Modeling Counterfactual Outcomes. Regardless of other choices, assigning individuals to the program with the highest estimated effectiveness reduces overall long-term unemployment and reemployment gaps (Figures 4 and 5). This represents gains due to explicit estimation of treatment effects rather than risk scores alone. For example: at baseline program sizes, when risk scores are not fairness constrained, targeting achieves a reduction of about 1.5 percentage points in overall long-term unemployment over random assignment, regardless of prioritization. If programs



(a) Optimal Program

(b) Random Program

Fig. 5. We plot overall long-term unemployment and the gender reemployment gap against program capacity for each combination of prioritization and assignment scheme. For clarity, results are shown only for fairness-unconstrained risk scores. Regardless of the assignment scheme, the Belgian prioritization results in slightly lower overall rates of long-term unemployment (blue line) and a smaller gender gap. Individualized program assignments (optimal) are markedly more effective.

are made five times larger, targeting achieves a reduction of about 3.7 percentage points over random assignment. Targeting is also much more effective than random assignment at reducing gender gaps under both prioritization regimes.

6 CONCLUSION AND FUTURE WORK

We have argued that algorithmic fairness requires anticipating the causal effects of deploying algorithms in concrete social settings on the distribution of outcomes. We have shown that existing methods in algorithmic fairness can have perverse distributive effects: requiring risk scores to be fair may exacerbate inequalities in social goods. Moreover, contrary to the accepted trade-offs between accurate and fair predictions, accurate prediction of individualized *counterfactual* outcomes supports policy in reducing inequality in the distribution of social goods.

Our approach has several limitations: we have not tried every fairness constraint, nor accounted for uncertainty in the estimation of individualized treatment effects. Realistic methods may have to make program assignments in an online, rather than a batch, fashion. Next to anticipatory evaluations, the design of algorithmically informed policies should also directly support the *ex-post* identification and evaluation of the policy. We have also adopted a rather paternalistic approach: future work should try to accommodate the preferences of the unemployed. Finally, we have simulated a policy approach that relies essentially on risk scores to facilitate prioritization. This reflects the state of algorithmic policy. However, risk scores increasingly seem like an unnecessary detour. We are inspired by the work of Körtner and Bach [51]: future work might directly seek distributively optimal allocations (perhaps with more sophisticated notions of optimality) without recourse to risk scores [42, 72]. This approach subjects claims of 'efficiency' to direct test and allows the conceptual innovations of theorists of distributive justice like Rawls and his interlocutors to flow directly into applications.

REFERENCES

[1] Ahmed Alaa, Zaid Ahmad, and Mark van der Laan. 2023. Conformal Meta-learners for Predictive Inference of Individual Treatment Effects. https://doi.org/10.48550/ARXIV.2308.14895

[2] Stefania Albanesi and Ayşegül Şahin. 2018. The gender unemployment gap. *Review of Economic Dynamics* 30 (2018), 47–67. https://doi.org/10.1016/j.red.2017.12.005

[3] Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data* 3 (feb 2020). https://doi.org/10.3389/fdata.2020.00005

[4] Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. *Der AMS-Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Endbericht.* Technical Report ITA-Projektbericht Nr. 2020-02. Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften. https://doi.org/10.1553/ita-pb-2020-02

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[6] Patrick Arni and Amelie Schiprowski. 2015. Die Rolle von Erwartungshaltungen in der Stellensuche und der RAV-Beratung - Teilprojekt 2: Pilotprojekt Jobchancen-Barometer. *IZA Research Reports* 70 (2015).

[7] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 62–76.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities.* MIT Press.

[9] Fabian Beigang. 2022. On the Advantages of Distinguishing Between Predictive and Allocative Fairness in Algorithmic Decision-Making. *Minds and Machines* 32, 4 (2022), 655–682.

[10] Richard A. Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. 2023. Fair Risk Algorithms. *Annual Review of Statistics and Its Application* 10, 1 (2023), 165–187. https://doi.org/10.1146/annurev-statistics-033021-120649

[11] Richard A Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. 2021. Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. *Sociological Methods & Research* (2021).

[12] Sebawit G. Bishu and Mohamad G. Alkadry. 2016. A Systematic Review of the Gender Pay Gap and Factors That Predict It. *Administration & Society* 49, 1 (2016), 65–104. https://doi.org/10.1177/0095399716636928

[13] Giuliano Bonoli. 2010. The Political Economy of Active Labor-Market Policy. *Politics & Society* 38, 4 (2010), 435–457. https://doi.org/10.1177/0032329210381235

[14] Denny Borsboom, Jan-Willem Romeijn, and Jelte M. Wicherts. 2008. Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods* 13, 2 (2008), 75–98. https://doi.org/10.1037/1082-989x.13.2.75

[15] Felix Bühlmann, Céline Schmid Botkine, Peter Farago, François Höpflinger, Dominique Joye, René Levy, Pasqualina Perrig-Chiello, and Christian Suter. 2013. *Swiss Social Report: Generations in Perspective.* Seismo. http://socialreport.ch/2012/first-level-page/long-term-unemployment/long-term-unemployment-in-switzerland-by-sex-1991-2010.html

[16] Bundesagentur für Arbeit. 2023. Statistik der Bundesagentur für Arbeit Berichte: Blickpunkt Arbeitsmarkt –Die Arbeitsmarktsituation von Frauen und Männern. *Nürnberg, May* (2023).

[17] Lucius E. J. Bynum, Joshua R. Loftus, and Julia Stoyanovich. 2022. Counterfactuals for the Future. https://doi.org/10.48550/ARXIV.2212.03974

[18] David Card, Jochen Kluve, and Andrea Weber. 2018. What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations. *Journal of the European Economic Association* 16, 3 (2018), 894–931. https://doi.org/10.1093/jeea/jvx028

[19] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (jan 2018), C1–C68. https://doi.org/10.1111/ectj.12097

[20] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[21] Bart Cockx, Michael Lechner, and Joost Bollens. 2023. Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium. *Labour Economics* 80 (2023), 102306. https://doi.org/10.1016/j.labeco.2022.102306

[22] Mark Considine, Phuc Nguyen, and Siobhan O'Sullivan. 2017. New public management and the rule of economic incentives: Australian welfare-to-work from job market signalling perspective. *Public Management Review* 20, 8 (2017), 1186–1204. https://doi.org/10.1080/14719037.2017.1346140

[23] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2018. The Measure and Mismeasure of Fairness. (2018). https://doi.org/10.48550/ARXIV.1808.00023

[24] Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* ACM. https://doi.org/10.1145/3351095.3372851

[25] Bruno Crépon, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment. *The Quarterly Journal of Economics* 128, 2 (2013), 531–580. https://doi.org/10.1093/qje/qjt001

[26] Alicia Curth, Richard W. Peck, Eoin McKinney, James Weatherall, and Mihaela van der Schaar. 2024. Using Machine Learning to Individualize Treatment Effect Estimation: Challenges and Opportunities. *Clinical Pharmacology & Therapeutics* (Jan. 2024). https://doi.org/10.1002/cpt.3159

[27] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. ACM, 525–534.

[28] S. Desiere, K. Langenbucher, and L. Struyven. 2019. Statistical profiling in public employment services. *OECD Social, Employment and Migration Working Papers* 224 (2019). https://doi.org/10.1787/b5e5f16e-en

[29] Sam Desiere and Ludo Struyven. 2020. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. *Journal of Social Policy* 50, 2 (2020), 367–385. https://doi.org/10.1017/s0047279420000203

[30] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM. https://doi.org/10.1145/2090236.2090255

[31] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway Feedback Loops in Predictive Policing. *Proceedings of Machine Learning Research* 81 (2018), 1–12.

[32] European Commission, Eurostat. Accessed 17 January 2024. *Long-term unemployment by sex - annual data.* https://ec.europa.eu/eurostat/databrowser/view/une_ltu_a/default/table?lang=en

[33] Daniel Goller, Tamara Harrer, Michael Lechner, and Joachim Wolff. 2021. Active labour market policies for the long-term unemployed: New evidence from causal machine learning. https://doi.org/10.48550/ARXIV.2106.10141

[34] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35, 4 (2022). https://doi.org/10.1007/s13347-022-00584-6

[35] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*. 3315–3323. https://doi.org/10.48550/ARXIV.1610.02413

[36] Lily Hu and Yiling Chen. 2018. A Short-term Intervention for Long-term Fairness in the Labor Market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. ACM Press. https://doi.org/10.1145/3178876.3186044

[37] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287600

[38] Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream Effects of Affirmative Action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287578

[39] Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3442188.3445919

[40] Christoph Kern, Ruben L. Bach, Hannah Mautner, and Frauke Kreuter. 2021. Fairness in Algorithmic Profiling: A German Case Study. https://doi.org/10.48550/ARXIV.2108.04134

[41] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[42] Toru Kitagawa and Aleksey Tetenov. 2019. Equality-Minded Treatment Choice. *Journal of Business & Economic Statistics* 39, 2 (2019), 561–574. https://doi.org/10.1080/07350015.2019.1688664

[43] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. https://doi.org/10.48550/ARXIV.1609.05807

[44] Henrik Kleven, Camille Landais, and Gabriel Leite-Mariante. 2023. The Child Penalty Atlas. *National Bureau of Economic Research* (2023). https://doi.org/10.3386/w31649

[45] Michael C. Knaus. 2022. Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal* 25, 3 (jun 2022), 602–627. https://doi.org/10.1093/ectj/utac015

[46] Michael C. Knaus, Michael Lechner, and Anthony Strittmatter. 2022. Heterogeneous Employment Effects of Job Search Programs: A Machine Learning Approach. *Journal of Human Resources* 57, 2 (2022), 597–636. https://doi.org/10.3368/jhr.57.2.0718-9615r1

[47] John Körtner and Giuliano Bonoli. 2021. Predictive Algorithms in the Delivery of Public Employment Services. *SocArXiv* (2021).

[48] Max Kunaschk and Julia Lang. 2022. Can Algorithms Reliably Predict Long-Term Unemployment in Times of Crisis? – Evidence from the COVID-19 Pandemic. *IAB-Discussion Paper* (2022). https://doi.org/10.48720/IAB.DP.2208

[49] Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. 2021. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There? https://doi.org/10.48550/ARXIV.2105.01441

[50] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[51] John Körtner and Ruben L. Bach. 2023. Inequality-Averse Outcome-Based Matching. (2023). https://doi.org/10.31219/osf.io/yrn4d

[52] Marloes Lammers and Lucy Kok. 2019. Are active labor market policies (cost-)effective in the long run? Evidence from the Netherlands. *Empirical Economics* 60, 4 (2019), 1719–1746. https://doi.org/10.1007/s00181-019-01812-3

[53] Michael Lechner. 1999. Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification. *Journal of Business & Economic Statistics* 17, 1 (Jan. 1999), 74–90. https://doi.org/10.1080/07350015.1999.10524798

[54] Michael Lechner, Michael Knaus, Martin Huber, Markus Frölich, Stefanie Behncke, Giovanni Mellace, and Anthony Strittmatter. 2020. Swiss Active Labor Market Policy Evaluation [Dataset]. *FORS, Lausanne, Switzerland* (2020).

[55] Karen Levy, Kyla E. Chasalow, and Sarah Riley. 2021. Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science* 17, 1 (2021), 309–334. https://doi.org/10.1146/annurev-lawsocsci-041221-023808

[56] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2019. Delayed Impact of Fair Machine Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/862

[57] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[58] Daniel Malinsky. 2018. Intervening on structure. *Synthese* 195, 5 (2018), 2295–2312.

[59] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 107–118. https://doi.org/10.48550/ARXIV.1705.09055

[60] Alan Mishler and Niccolò Dalmasso. 2022. Fair When Trained, Unfair When Deployed: Observable Fairness Measures are Unstable in Performative Prediction Settings.

[61] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[62] Andreas Mueller and Johannes Spinnewijn. 2023. The Nature of Long-Term Unemployment: Predictability, Heterogeneity and Selection. *National Bureau of Economic Research* (2023). https://doi.org/10.3386/w30979

[63] Nicolò Pagan, Joachim Baumann, Ezzat Elokda, Giulia De Pasquale, Saverio Bolognani, and Anikó Hannák. 2023. A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM. https://doi.org/10.1145/3617694.3623227

[64] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.

[65] Anette C. M. Petersen, Lars Rune Christensen, Richard Harper, and Thomas Hildebrandt. 2021. "We Would Never Write That Down": Classifications of Unemployed and Data Challenges for AI. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26. https://doi.org/10.1145/3449176

[66] Sebastian Scher, Simone Kopeinik, Andreas Trügler, and Dominik Kowald. 2023. Modelling the long-term fairness dynamics of data-driven targeted help on job seekers. *Scientific Reports* 13, 1 (2023).

[67] Marco Scutari, Francesca Panero, and Manuel Proissl. 2022. Achieving fairness with a simple ridge penalty. *Statistics and Computing* 32, 5 (Sept. 2022). https://doi.org/10.1007/s11222-022-10143-w

[68] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287598

[69] Eran Tal. 2023. Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. ACM. https://doi.org/10.1145/3600211.3604678

[70] Eran Tal. 2023. Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 312–321.

[71] Alexander Williams Tolbert and Emily Diana. 2023. Correcting Underrepresentation and Intersectional Bias for Fair Classification. https://doi.org/10.48550/ARXIV.2306.11112

[72] Davide Viviano and Jelena Bradic. 2023. Fair Policy Targeting. *J. Amer. Statist. Assoc.* (2023), 1–14. https://doi.org/10.1080/01621459.2022.2142591

[73] Melvin Vooren, Carla Haelermans, Wim Groot, and Henriëtte Maassen van den Brink. 2018. The Effectivnesness of Active Labor Market Policies: A Meta-Analysis. *Journal of Economic Surveys* 33, 1 (2018), 125–149. https://doi.org/10.1111/joes.12269

[74] Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. 2023. Algorithmic Unfairness through the Lens of EU Non-Discrimination Law. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3593013.3594044

[75] Xueru Zhang and Mingyan Liu. 2021. Fairness in Learning-Based Sequential Decision Algorithms: A Survey. In *Handbook of Reinforcement Learning and Control*. Springer International Publishing, 525–555. https://doi.org/10.1007/978-3-030-60990-0_18

[76] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems* (2020), 1–13.

A PROOF OF THEOREM 4.1

Proof of Theorem 4.1. First, we need to show that all terms are well-defined. This amounts to showing that $P_{\text{post}}(A = a, X = x)$, $P_{\text{pre}}(A = a)$ and $P_{\text{pre}}(A = a, X = x, D = d)$ are strictly greater than zero for all $(x, d) \in \Pi_{\text{post}}$.

We first show that $P_{\text{pre}}(A = a) > 0$. Note that

$$\begin{aligned} P_{\text{pre}}(A = a) &= \sum_{x \in \mathcal{X}} P_{\text{pre}}(A = a, X = x) \\ &= \sum_{x \in \mathcal{X}} P_{\text{post}}(A = a, X = x) && \text{(No Feedback)} \\ &= P_{\text{post}}(A = a) > 0. \end{aligned}$$

We now show that $P_{\text{post}}(A = a, X = x) > 0$ for all $(x, d) \in \Pi_{\text{post}}$. Note that

$$\begin{aligned} P_{\text{post}}(A = a, X = x) &= P_{\text{post}}(A = a) \sum_{e \in \mathcal{D}} P_{\text{post}}(X = x, D = e | A = a) \\ &\geq P_{\text{post}}(A = a) P_{\text{post}}(X = x, D = d | A = a) > 0. \end{aligned}$$

Finally, we show that $P_{\text{pre}}(A = a, X = x, D = d) > 0$ for all $(x, d) \in \Pi_{\text{post}}$. Since $P_{\text{pre}}(A = a) > 0$, it suffices to show that $P_{\text{pre}}(X = x, D = d | A = a) > 0$ for all $(x, d) \in \Pi_{\text{post}}$. Accordingly, suppose that $(x, d) \in \Pi_{\text{post}}$. Then

$$P_{\text{post}}(X = x, D = d | A = a) = P_{\text{post}}(D = d | X = x, A = a) P_{\text{post}}(X = x | A = a) > 0,$$

which entails that both $P_{\text{post}}(D = d | X = x, A = a) > 0$ and $P_{\text{post}}(X = x | A = a) > 0$. By No Unprecedented Decisions, $P_{\text{pre}}(D = x | X = x, A = a) > 0$ and by No Feedback $P_{\text{pre}}(X = x | A = a) > 0$. Therefore,

$$P_{\text{pre}}(X = x, D = d | A = a) = P_{\text{pre}}(D = x | X = x, A = a) P_{\text{pre}}(X = x | A = a) > 0;$$

and the question of well-definedness is settled.

Next, note that: $P_{\text{post}}(Y = y \mid A = a) =$

$$\begin{aligned} &= \sum_{(x,d) \in \Pi_{\text{post}}} P_{\text{post}}(Y = y \mid A = a, X = x, D = d) P_{\text{post}}(X = x, D = d \mid A = a) && \text{(Total Probability)} \\ &= \sum_{(x,d) \in \Pi_{\text{post}}} P_{\text{post}}(Y = y \mid A = a, X = x, D = d) P_{\text{post}}(X = x | A = a) P_{\text{post}}(D = d \mid A = a, X = x) \\ &= \sum_{(x,d) \in \Pi_{\text{post}}} P_{\text{post}}(Y = y \mid A = a, X = x, D = d) P_{\text{pre}}(X = x | A = a) P_{\text{post}}(D = d \mid A = a, X = x). && \text{(No Feedback)} \end{aligned}$$

Note that, whenever defined,

$$\begin{aligned}
P_t(Y=y \mid A=a, X=x, D=d) &= P_t\left(\sum_{e\in\mathcal{D}} Y^e \mathbb{1}[D=e]=1 \mid A=a, X=x, D=d\right) && \text{(Consistency)}\\
&= P_t\left(Y^d = y \mid A=a, X=x, D=d\right)\\
&= P_t\left(Y^d = y \mid A=a, X=x\right). && \text{(Unconfoundedness)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
P_{\text{post}}(Y=y \mid A=a, X=x, D=d) &= P_{\text{post}}\left(Y^d = y \mid A=a, X=x\right)\\
&= P_{\text{pre}}\left(Y^d = y \mid A=a, X=x\right) && \text{(Stable CATE)}\\
&= P_{\text{pre}}(Y=y \mid A=a, X=x, D=d);
\end{aligned}$$

and therefore $P_{\text{post}}(Y=y \mid A=a) =$

$$= \sum_{(x,d)\in\Pi_{\text{post}}} P_{\text{pre}}(Y=y \mid A=a, X=x, D=d) P_{\text{pre}}(X=x|A=a) P_{\text{post}}(D=d \mid A=a, X=x).$$

□

B CASE STUDY

B.1 Descriptive Statistics: Simulation Data

	#Obs	LTU	Female (binary)	Age in years	Non-Citizen (binary)	Employability	Past Income in CHF
Simulation Data	32,148	0.41	0.44	36.8	0.36	1.93	43,461
No program	23,785	0.41	0.43	36.6	0.37	1.92	42,557
Vocational	423	0.28	0.32	37.5	0.32	1.91	49,349
Computer	446	0.24	0.61	38.9	0.20	1.98	43,251
Language	723	0.48	0.54	35.3	0.68	1.83	37,779
Job Search	5,868	0.43	0.44	37.4	0.33	1.98	46,815
Employment	321	0.46	0.43	35.3	0.39	1.84	36,902
Personality	582	0.37	0.35	39.4	0.25	1.93	53,136

Table 1. Descriptive statistics for key demographic variables in the test and simulation data and by observed treatment groups. Long-term unemployment (LTU), Female, and Non-Citizen are given as shares. Age, Employability, and Past Income are averages. Employability is an ordered variable from low (1) to high (3), assigned by the caseworker. Knaus [45] reports an exchange rate USD/CHF of about 1.3 for 2003.

B.2 Descriptive Statistics: Full sample

	#Obs	LTU	Female (binary)	Age in years	Non-Citizen (binary)	Employability	Past Income in CHF
Full Sample	64,296	0.41	0.44	36.8	0.36	1.93	43,391
No program	47,631	0.41	0.44	36.6	0.37	1.93	42,529
Vocational	858	0.29	0.33	37.5	0.30	1.93	48,654
Computer	905	0.28	0.60	39.1	0.21	1.97	43,213
Language	1,504	0.47	0.55	35.28	0.66	1.85	37,300
Job Search	11,610	0.43	0.44	37.3	0.33	1.98	46,693
Employment	611	0.43	0.41	35.3	0.38	1.83	37,084
Personality	1,177	0.37	0.36	38.7	0.27	1.93	53,067

Table 2. Descriptive statistics for key demographic variables in the full sample and by observed treatment groups. The simulation data is drawn from this full sample. Long-term unemployment (LTU), Female, and Non-Citizen are given as shares. Age, Employability, and Past Income are averages. Employability is an ordered variable from low (1) to high (3), assigned by the caseworker. Knaus [45] reports an exchange rate USD/CHF of about 1.3 for 2003.

B.3 Double-Robust Machine Learning for Estimating IAPOs

In Section 4, we have theoretically derived conditions under which the post-interventional gender gap is identified. Two assumptions concern the internal validity of our study. Unconfoundedness is the strongest assumption. Replicating the work by Knaus [45], Knaus et al. [46] and Körtner and Bach [51], we rely on extensive information on caseworkers and their subjective assessment of their clients in the estimation of treatment effects combined with rich administrative data on the demographics and employment biographies to support the assumption. No Unprecedented decisions requires that the propensity scores are non-zero. The other two concern the external validity of our simulation study. We presuppose that the treatment effects of the programs are stable under different allocations and increased program capacities (stable CATEs) and that the pool of unemployed stays the same (No Feedback on the covariates).

First, we estimate the normalized conditional probability to be allocated into each program (the propensity of treatment, $e_d(X_i)$) and the conditional outcome mean in the observed allocation (in short, conditional outcome, $\mu(d,x)$). Given the small number of observations in most of the labor market programs, we use the full data set and cross-validation for the estimation of the nuisance parameters. The two nuisance parameters then allow the estimation of the doubly robust score:

$$\hat{\Gamma}_{i,d} = \hat{\mu}(d, X_i) + \frac{D_i(d)(Y_i - \hat{\mu}(d, X_i))}{\hat{e}_d(X_i)},$$

where $D_i(d)$ indicates the treatment assignment for individual i and Y_i the observed, pre-interventional outcome. This strategy is called doubly robust because the functional form of either the propensity score or the conditional outcome can be miss-specified without threatening the identification [19, 45]. In the last step, the estimates of the debiased scores, $\hat{\Gamma}_{i,d}$, are used as pseudo outcomes to estimate the conditional expected outcomes, $\mathrm{E}[\hat{\Gamma}_{i,d} \mid X_i]$ using a regression forest. These estimates are the individualized average potential outcomes for each treatment option under the outlined identifying assumptions. For this step, the regression forest is trained only on the training set.

We estimate individualized average treatment effects for each individual i in the sample as differences between the respective individualized average potential outcomes:

$$\hat{\Delta}_{i,d,d'} = \hat{\Gamma}_{i,d} - \hat{\Gamma}_{i,d'}.$$

In Table 6, we show the distribution of individualized average treatment effects by gender. While the overall trends remain the same, all treatments except job search programs on average are slightly more effective for women than for men. Treatment effects are estimated against the baseline of no program.

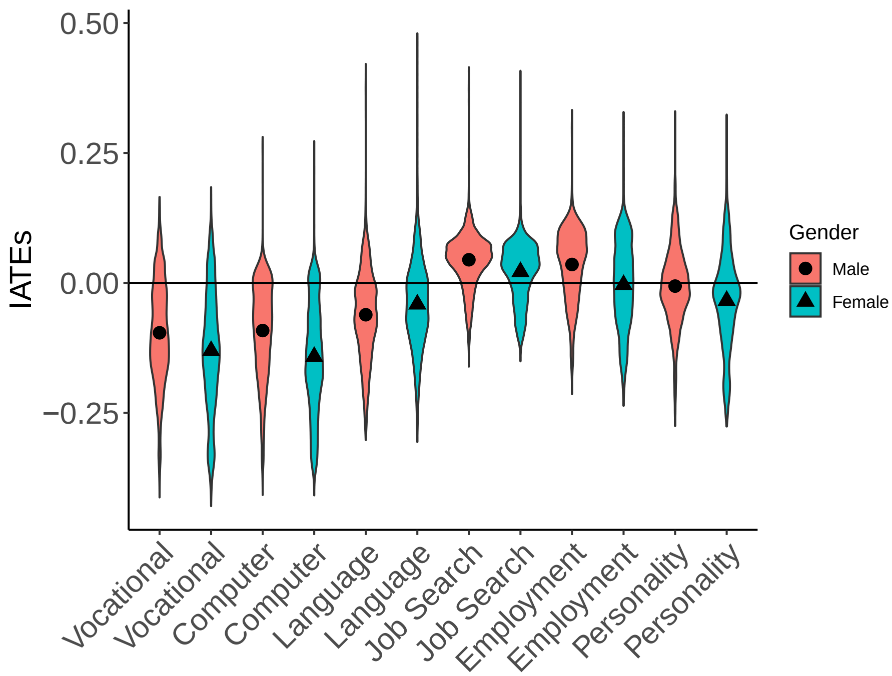


Fig. 6. Individualized and Average Treatment Effects for all six labor market programs by gender. Baseline is "no program".

B.4 Risk Scores and Prioritization Policies

To determine the prioritization of registered unemployed in its Belgian or Austrian variants we estimate risk scores for becoming long-term unemployed. The full list of features is given in Table 3. For a discussion on the predictability of long-term unemployment, see Mueller and Spinnewijn [62]. Using administrative data from Germany, Kunaschk and Lang [48] evaluate the performance of risk scores under external shocks like the COVID-19 pandemic. Kern et al. [40] evaluate the violation of retrospective fairness criteria when predicting long-term unemployment in the same context.

First, we estimate risk scores by a fairness-unconstrained logistic ridge regression. The optimal regularization strength is chosen by cross-validation at about $\lambda = 0.049$. Second, we add a fairness constraint for *statistical parity* and, third, a constraint for *equal opportunity*. In this case, the true positive rates among the sensitive attribute are equalized, a relaxation of Separation [35]. We make use of the the implementation by [67] for the estimation of fairness constrained risk scores. To achieve statistical parity they use a ridge penalty to bound the variance explained by the sensitive attribute (gender) over the total explained variance. For equal opportunity, the risk score is regressed against the sensitive attribute and the outcome variable with the ridge penalty bounding the variance explained by the sensitive attribute over the total explained variance. In both cases, we use a fairness penalty of 0.01, where 0 requires perfect fairness and 1 corresponds to no fairness constraint.

Note some important differences between the Belgian and Austrian implementations of our work. In Flanders, Belgium the probability of re-employment within six months is estimated by a random forest model [29]. Sensitive attributes are no longer included due to privacy regulations. In our simulation study, the definition of long-term unemployment corresponds to the ILO definition with 12 months of uninterrupted unemployment.

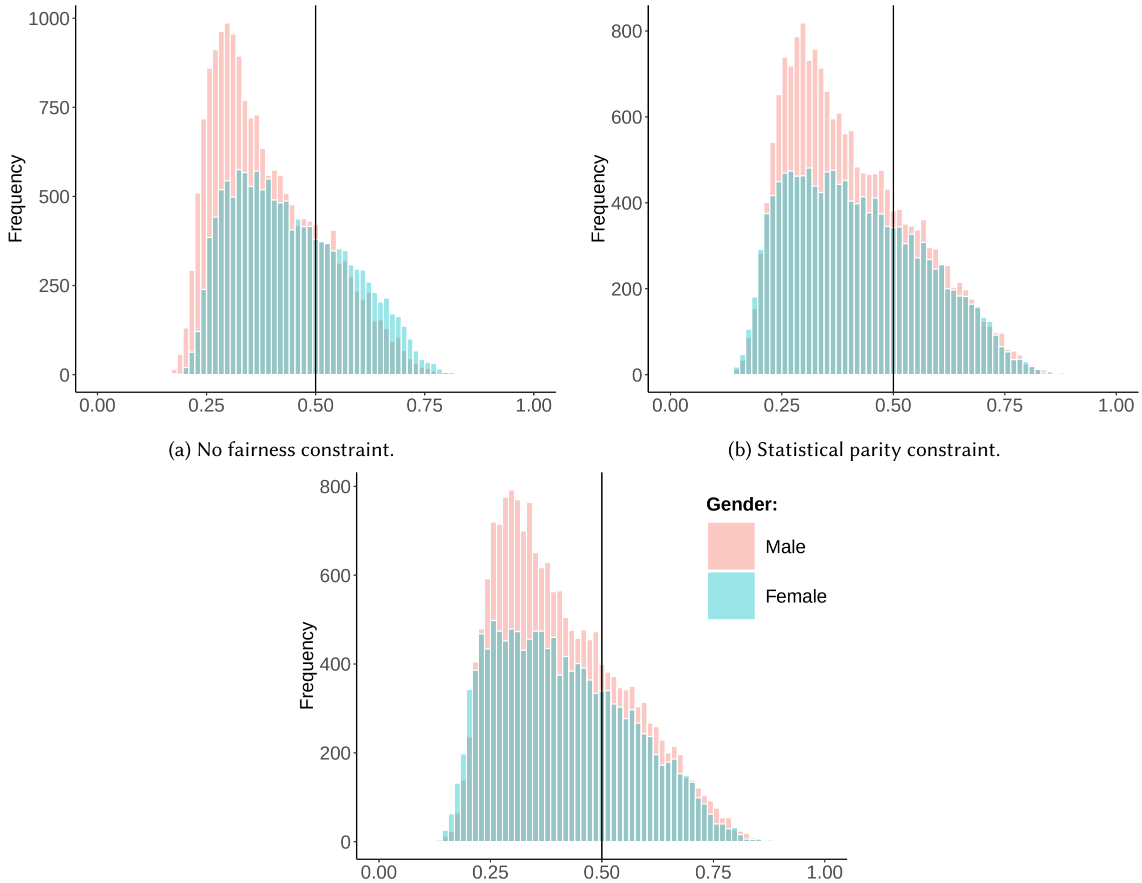
In Austria, two different models are estimated [4]. The first, short-term model, uses as a binary target at least 90 days of unsupported employment within seven months after the reference date. The second, long-term model, uses at least 180 days of unsupported employment within 24 months as the target. Those with a short-term probability of employment of above 66% are classified as low risk for LTU. Those with a long-term probability of employment below 25% are classified as high risk. The middle group is built as a residual. That is, it includes all those not classified as high or low risk. In difference to earlier reports [3], a stratification approach is applied, and logistic regressions are used to evaluate the feature importance only [4]. Sensitive attributes like gender and citizenship are included as features. In difference to the Austrian proposal, we estimate one model and create the prioritized middle group as those individuals falling in the $30 - 70th$ percentile of the respective risk distribution.

Features for the estimation of risk scores
Age
Mother tongue in canton's language
Lives in big city
Lives in medium city
Lives in no city
Fraction of months employed in last 2 years
Number of employment spells in last 5 years
Female (binary)
Foreigner with temporary permit
Foreigner with permanent permit
Cantonal GDP p.c.
Married
Mother tongue other than German, French, Italian
Past income in CHF
Previous job: Manager
Previous job in missing sector
Previous job in primary sector
Previous job in secondary sector
Previous job in tertiary sector
Previous job: self-employed
Previous job: skilled worker
Previous job: unskilled worker
Qualification: semiskilled
Qualification: some degree
Qualification: unskilled
Qualification: skilled without degree
Swiss citizenship
Number of unemployment spells in last 2 years
Cantonal unemployment rate in %

Table 3. List of features used for the estimation of risk scores. All caseworker information is omitted from this estimation. See Lechner et al. [54] for detailed information about the administrative data.

	Reference	Ridge Regression	Statistical Parity	Equality of Opportunity
Accuracy	(1)	0.644	0.644	0.645
Precision	(1)	0.612	0.605	0.607
Recall	(1)	0.384	0.404	0.404
Stat Parity	(0)	**0.116**	**0.041**	0.019
Equal Opportunity	(0)	**0.173**	0.07	**0.044**
False Positive Parity	(0)	**0.062**	0.005	-0.014
Positive Predictive Parity	(0)	0.062	0.072	0.081
Negative Predictive Parity	(0)	0.011	0.011	0.016

Table 4. Results for predicting long-term unemployment. To achieve predictions of the binary target variable, a threshold of 0.5 is applied to the risk scores.



(a) No fairness constraint.

(b) Statistical parity constraint.

(c) Equal opportunity constraint.

Fig. 7. Risk scores estimated by logistic ridge regression, with and without fairness constraints. The vertical line at .5 gives the decision threshold for binary predictions.

B.5 Results from the Simulation Study

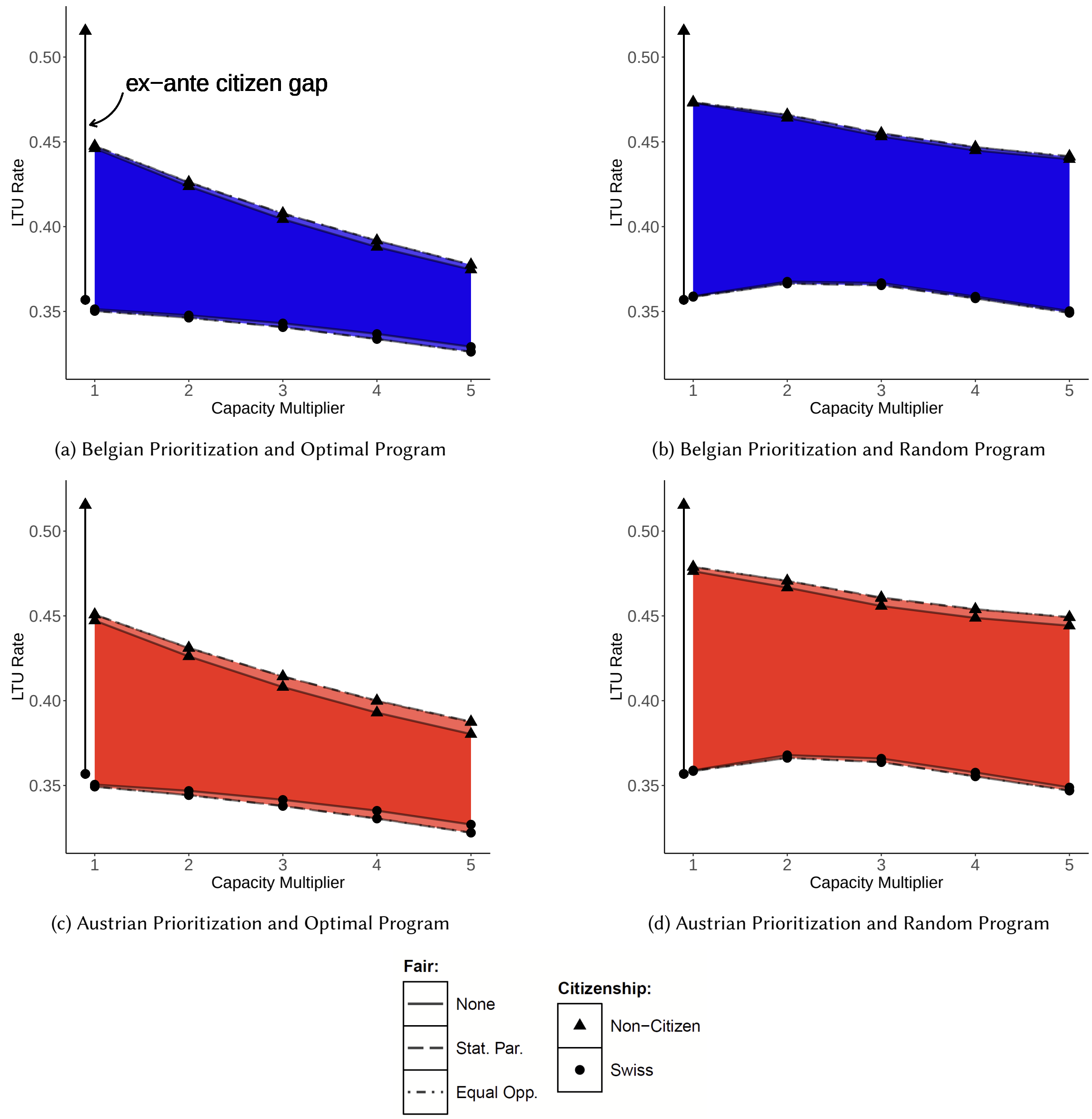
	LTU	Women	Men	Gender gap	Non-Citizens	Citizen	Citizen Gap
Status quo	0.414	0.436	0.397	0.039	0.515	0.357	0.158
Belgian, optimal							
Logistic Regression	0.386	0.404	0.372	0.032	0.446	0.351	0.095
Stat. Parity	0.386	0.408	0.368	0.039	0.448	0.35	0.097
Equal Opp.	0.386	0.409	0.368	**0.041**	0.448	0.35	0.097
Belgian, random							
Logistic Regression	0.4	0.421	0.385	0.036	0.473	0.359	0.114
Stat. Parity	0.400	0.422	0.383	0.039	0.473	0.359	0.114
Equal Opp.	0.400	0.423	0.383	**0.04**	0.474	0.359	0.115
Austrian, optimal							
Logistic Regression	0.386	0.405	0.371	0.034	0.447	0.351	0.097
Stat. Parity	0.386	0.408	0.369	0.038	0.451	0.349	0.101
Equal Opp.	0.386	0.408	0.369	0.039	0.451	0.349	0.101
Austrian, random							
Logistic Regression	0.402	0.423	0.385	0.037	0.476	0.359	0.117
Stat. Parity	0.402	0.424	0.385	**0.04**	0.479	0.358	0.120
Equal Opp.	0.402	0.425	0.385	**0.04**	0.479	0.359	0.120

Table 5. Rates of in long-term unemployment for the different algorithmically informed policies under **baseline** capacities.

	LTU	Women	Men	Gender Gap	Citizens	Non-Citizen	Citizen Gap
Status quo	0.414	0.436	0.397	0.039	0.515	0.357	0.158
Belgian, optimal							
Logistic Regression	0.346	0.351	0.342	0.009	0.375	0.329	0.046
Stat. Parity	0.345	0.36	0.333	0.026	0.378	0.326	0.051
Equal Opp.	0.345	0.362	0.332	0.03	0.377	0.326	0.051
Belgian, random							
Logistic Regression	0.383	0.395	0.373	0.022	0.44	0.350	0.09
Stat. Parity	0.383	0.399	0.370	0.029	0.441	0.349	0.092
Equal Opp.	0.383	0.400	0.369	0.031	0.442	0.349	0.092
Austrian, optimal							
Logistic Regression	0.346	0.353	0.341	0.012	0.380	0.327	0.053
Stat. Parity	0.346	0.361	0.334	0.026	0.388	0.322	0.066
Equal Opp.	0.346	0.362	0.333	0.029	0.387	0.322	0.065
Austrian, random							
Logistic Regression	0.383	0.397	0.373	0.024	0.444	0.349	0.095
Stat. Parity	0.384	0.401	0.371	0.030	0.449	0.347	0.102
Equal Opp.	0.384	0.402	0.370	0.032	0.449	0.347	0.102

Table 6. Rates in long-term unemployment for the different algorithmically informed policies under **five-fold** capacities.

B.6 Citizen LTU Gap



(a) Belgian Prioritization and Optimal Program

(b) Belgian Prioritization and Random Program

(c) Austrian Prioritization and Optimal Program

(d) Austrian Prioritization and Random Program

Fig. 8. We plot the citizen gap in long-term unemployment (LTU) against program capacity for each combination of prioritization and assignment scheme. The level of transparency shows the citizen gap for the corresponding fairness constraint: none, statistical parity, or equal opportunity. All policy combinations reduce the citizen gap. The unconstrained risk scores (lowest transparency) result in the smallest citizen gap. This effect is especially pronounced as program capacity is increased and program assignments are individualized (optimal). Austrian prioritization compared to the Belgian approach performs particularly poorly under fairness constraints for gender.

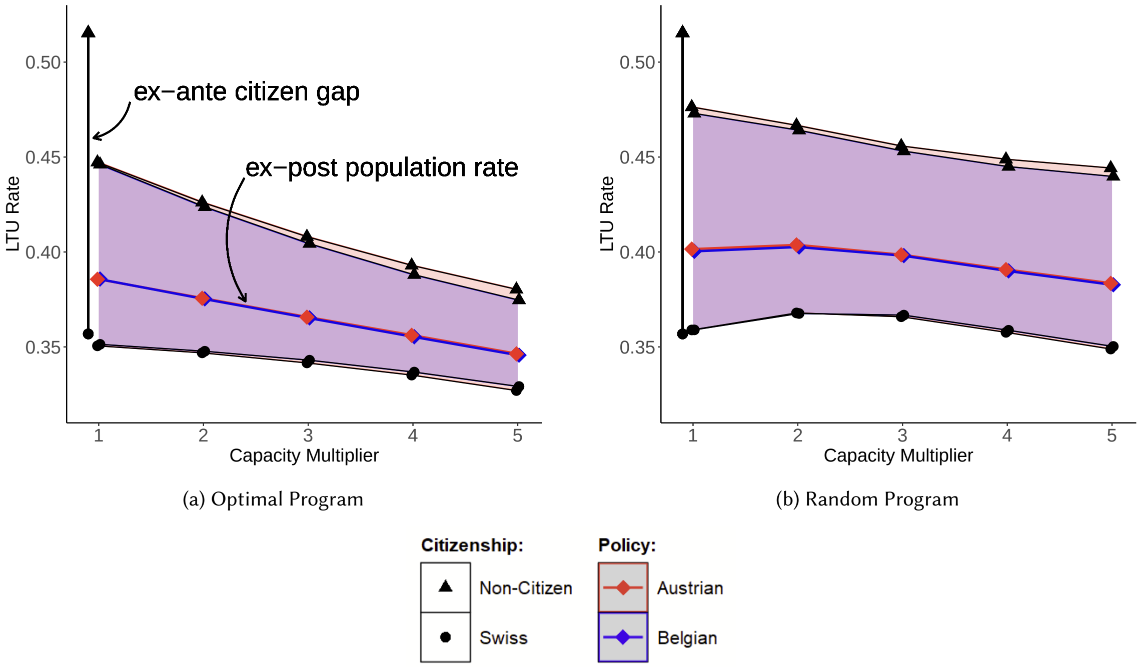
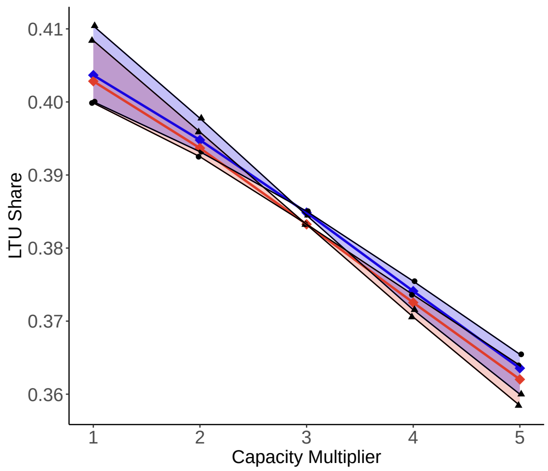
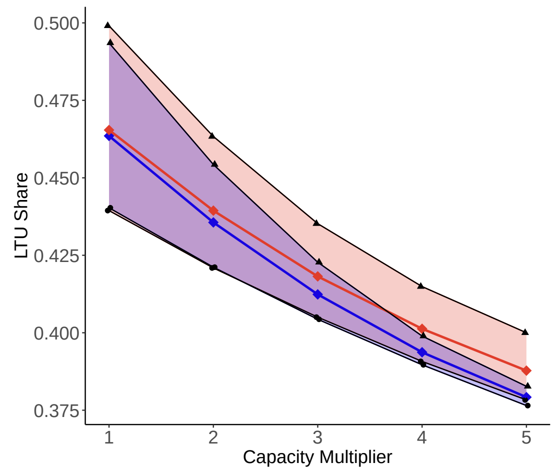


Fig. 9. We plot overall long-term unemployment and the citizen reemployment gap against program capacity for each combination of prioritization and assignment scheme. For clarity, results are shown only for fairness-unconstrained risk scores. Regardless of the assignment scheme, the Belgian prioritization (blue line) results in the same long-term unemployment rate as the Austrian and a slightly smaller citizen gap. Individualized program assignments (optimal) are markedly more effective, especially under larger program capacities.

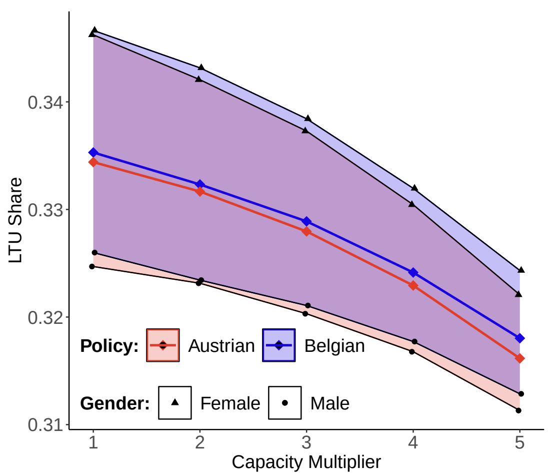
B.7 Gender LTU Gaps for (un)married (non-)citizen



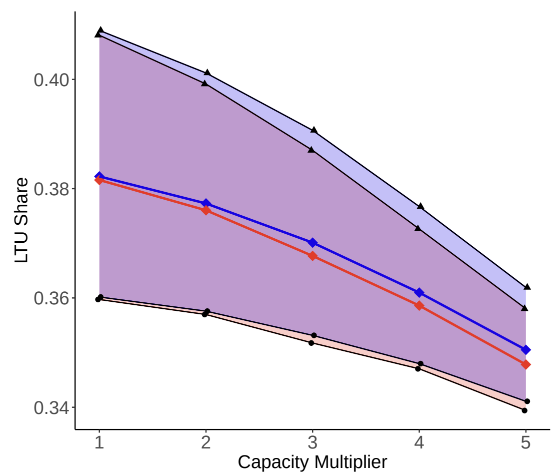
(a) Unmarried Non-Citizen



(b) Married Non-Citizen



(c) Unmarried Swiss Citizen



(d) Married Swiss Citizen

Fig. 10. Long-term unemployment rates among the respective group (red and blue line) and by gender for four sub-groups: unmarried non-citizen, unmarried Swiss citizen, married non-citizen, and married Swiss citizen. Note the different scales. The reduction in LTU rates and the gender gap is especially pronounced for the group of married foreigners. For unmarried foreigners, the gender gap even flips under both algorithmic policies at four- and five-fold program capacities.