Performativity and Prospective Fairness

Sebastian Zezulka
University of Tübingen
sebastian.zezulka@uni-tuebingen.de

Konstantin Genin
University of Tübingen
konstantin.genin@uni-tuebingen.de

Abstract

Deploying an algorithmically informed policy is a significant intervention in the structure of society. As is increasingly acknowledged, predictive algorithms have performative effects: using them can shift the distribution of social outcomes away from the one on which the algorithms were trained. Algorithmic fairness research is usually *motivated* by the worry that these performative effects will exacerbate the structural inequalities that gave rise to the training data. However, standard retrospective fairness methodologies are ill-suited to predict these effects. They impose static fairness constraints that hold after the predictive algorithm is trained, but before it is deployed and, therefore, before performative effects have had a chance to kick in. However, satisfying static fairness criteria after training is not sufficient to avoid exacerbating inequality after deployment. Addressing the fundamental worry that motivates algorithmic fairness requires explicitly comparing the change in relevant structural inequalities *before* and *after* deployment. We propose a prospective methodology for estimating this post-deployment change from pre-deployment data and knowledge about the algorithmic policy. That requires a strategy for distinguishing between, and accounting for, different kinds of performative effects. In this paper, we focus on the algorithmic effect on the causally downstream outcome variable. Throughout, we are guided by an application from public administration: the use of algorithms to (1) predict who among the recently unemployed will stay unemployed for the long term and (2) targeting them with labor market programs. We illustrate our proposal by showing how to predict whether such policies will exacerbate gender inequalities in the labor market.

1 A fundamental question for fair machine learning

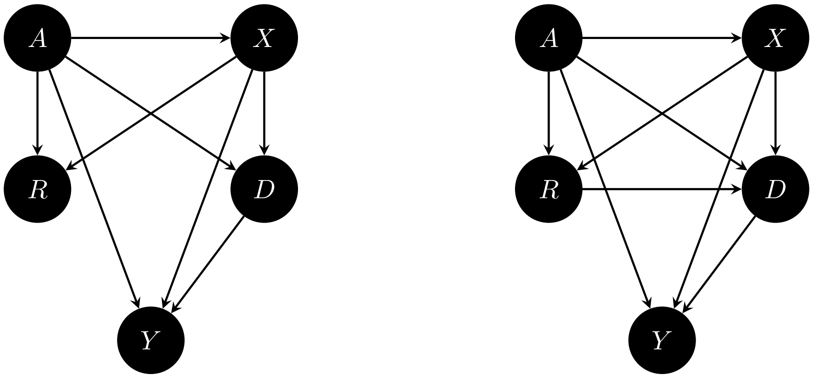
Research in algorithmic fairness is usually motivated by the worry that machine learning algorithms will reproduce or exacerbate the structural inequalities in the social processes that gave rise to their training data [Lum and Isaac, 2016, Tolbert and Diana, 2023]. Indeed, whether or not an algorithm exacerbates an existing social inequality is emerging as a central compliance criterion in EU non-discrimination law [Weerts et al., 2023]. However, the methodological solutions developed by researchers in algorithmic fairness are, surprisingly, ill-suited for answering this fundamental question. In order to decide whether embedding some algorithm into our socio-technical processes exacerbates existing structural inequalities, we must make some effort to, first, identify the contextually relevant inequalities and, second, predict whether the new algorithmic policy will exacerbate them. Most algorithmic fairness methods are *retrospective* in so far as they usually do not attempt the latter. Moreover, by failing to have this latter goal in mind, they typically also struggle to identify the contextually relevant inequalities, focusing instead on internal features of the algorithm. Since most structural inequalities long predate prediction algorithms, internal fairness properties of the algorithm are, at best, a proxy for the relevant structural inequality.

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

Preprint. Under review.

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 case-worker 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 [Barocas et al., 2019] ignores all but the first parameter; causal fairness [Kilbertus et al., 2017, Kusner et al., 2017] ignores the last; and individual fairness [Dwork et al., 2012] 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.* Our first point is that addressing the fundamental question of fair machine learning is a matter of 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.



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

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

Figure 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 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 $\varphi(G_{\text{post}}, P_{\text{post}}, M)$.

Deploying a machine learning algorithm, as shown in Figure 1, introduces an additional causal path from the predicted risk scores, R, to the decisions made, D. Importantly, the outcome variable Y is causally downstream of this intervention. In line with Malinsky [2018] and Bynum et al. [2022], we use an expanded notion of structural intervention that allows for the introduction of new causal paths, and not just the removal of existing paths as in standard atomic $do(X = x)$ interventions that fix a variable to a specific value.

From the 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 [2022] show that satisfying the fairness notions of sufficiency or separation[1] at the time of training virtually ensures that they will be violated after deployment. Illustrating the point in terms of sufficiency, where $\perp\!\!\!\perp$ denotes (conditional) statistical independence:

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

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 [Perdomo et al., 2020]. But as Mishler and Dalmasso [2022] 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 will change after deployment, Berk et al. [2021] advises against group-based metrics involving the outcome Y, opting for simple independence ($R \perp\!\!\!\perp A$) instead. Of course, independence requires non-trivial losses in predictive accuracy, which could undermine the effectiveness of 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. [2017] 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 *probative* of whether the algorithm reproduces or exacerbates existing inequalities since these effects are causally *downstream* of algorithmic predictions. In particular, it is customary to ignore the dependence between A and Y induced by the social status quo, 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 [Liu et al., 2019, Zhang and Liu, 2021]. Streamlined dynamical models and simulations are a valuable tool in evaluating the long-run effects of fairness-constrained algorithms. The main contribution of this paper, however, is a methodology for estimating the systemic effect of a proposed algorithmic policy from pre-deployment data. Of course, we would not expect such a procedure to exist in all cases. Rather, we hope to show that this is possible under not-too-heroic assumptions.

The plan of the paper is as follows: in the next section we survey related work; section 2 introduces the variety of algorithmic policies that have been proposed to support public employment agencies in reducing long-term unemployment; we argue that, in this context, the dependence between gender and employment outcome, as well as the gender gap in reemployment probabilities, are simple and intuitive measures $\varphi(\cdot)$ of systemic inequality; section 3 enumerates the challenges posed by different performative effects for predicting a post-deployment measure of systemic inequality from pre-deployment data and propose a method for overcoming (some of) them; section 4 illustrates the method with a simple example and section 5 outlines directions for future work.

1.1 Related Work

The fairness debate in machine learning began with debates about risk assessment tools for decision- and policy-making [Angwin et al., 2016, Kleinberg et al., 2016, Chouldechova, 2017, Mitchell et al., 2021]. To this day, many standard case studies e.g., lending, school admissions, and pretrial detention, fall within this scope. See Berk et al. [2023] for a review on fairness in risk assessment.[2] Since

[1] Respectively, that $Y \perp\!\!\!\perp A \mid R$ and $R \perp\!\!\!\perp A \mid Y$ [Barocas et al., 2019].

[2] Predecessors of this debate can be found in the psychometric literature, see Borsboom et al. [2008] and Hutchinson and Mitchell [2019].

then, researchers have stressed the importance of explicitly differentiating policy decisions from the risk predictions that inform them [Barabas et al., 2018, Kuppler et al., 2021, Beigang, 2022] and of studying machine learning algorithms in their socio-technological contexts [Selbst et al., 2019]. 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 [2022] show that a classifier that is formally fair in the training distribution will violate the respective fairness constraint in the post-deployment distribution. For this reason, Berk et al. [2021] argues for independence (demographic parity) as a fairness constraint, because it does not feature the outcome variable Y. Coston et al. [2020] 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 probative of the policy's effect on structural inequality. This paper's main contribution is to build upon the negative results of Berk et al. [2021] and Mishler and Dalmasso [2022]: we show how the post-interventional effect of an algorithmically-informed policy on a structural inequality can be identified from a combination of (1) observational, pre-deployment data and (2) knowledge 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 [2021] for a survey. Ensign et al. [2018] consider predictive feedback loops from selective data collection in predictive policing. Hu and Chen [2018] propose short-term interventions in the labor market to achieve long-term objectives. Using two-stage models, Liu et al. [2019] and Kannan et al. [2019] show that procedural fairness constraints can, under some conditions, have negative effects on outcomes in disadvantaged groups. D'Amour et al. [2020] confirm with simulation studies that imposing static fairness constraints does not guarantee that these constraints are met over time and can, under some conditions, exacerbate structural inequalities. Similar work is done by Zhang et al. [2020]. Creager et al. [2020] propose to unify dynamical fairness approaches in a causal DAG framework. Using time-lagged graphs, Hu and Zhang [2022] formulate a version of counterfactual long-term fairness. 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 procedural fairness notions designed to hold in the training distribution.

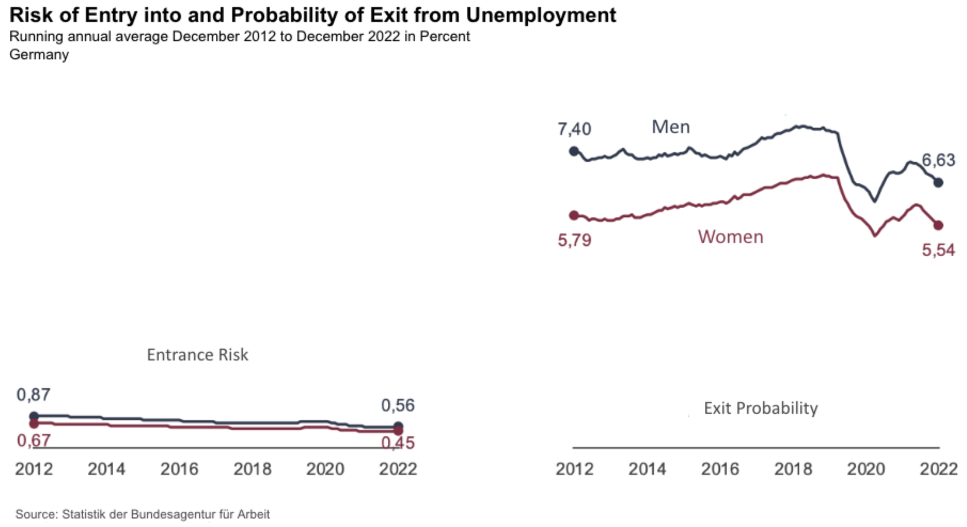
In this paper, we focus on using statistical profiling by public employment services to allocate the recently unemployed into active labor market programs. Respectively, Desiere and Struyven [2020] and Allhutter et al. [2020] provide detailed studies of the existing Flemish, and the proposed Austrian, algorithms. Using administrative data, Kern et al. [2021] perform a hypothetical analysis in a German setting. Scher et al. [2023] propose a dynamical model to study long-term and feedback effects on skills in a labor market context. To reduce inequality in the outcome distribution, Körtner and Bach [2023] propose an inequality-averse objective function for the allocation of people into labor market programs. Kitagawa and Tetenov [2019] and Viviano and Bradic [2023] make similar proposals in a more general setting.

2 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 [Considine et al., 2017]. ALMPs take many forms, but paradigmatic examples include resume workshops, job-training programs and placement services, see Bonoli [2010] for a helpful taxonomy. Evaluations of ALMPs across OECD countries find small but positive effects on labor market outcomes [Card et al., 2018, Vooren et al., 2018, Lammers and Kok, 2019]. 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 [Knaus et al., 2020], Belgium [Cockx et al., 2023], and Germany [Goller et al., 2021]. This implies potential welfare gains from a more targeted allocation into programs, especially when taking into account opportunity costs. This is one compelling motivation for the algorithmic support of allocation decisions.

Statistical profiling of the unemployed is now current practice in various OECD countries including Australia, the Netherlands and Flanders, Belgium [Desiere et al., 2019]. Paradigmatically, supervised learning techniques are employed to predict who is at risk of becoming long-term unemployed (LTU).[3] These tools are regularly framed as introducing objectivity and effectiveness in the provision of public goods and align with demands for evidence-based policy and digitisation in public administration.[4]

Individual scores predicting the risk of long-term unemployment support a variety of decisions. For example, the public employment service (PES) of Flanders uses risk scores so far only to help caseworkers and line managers decide who to contact first, prioritizing those at higher risk [Desiere and Struyven, 2020]. 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 first 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 [Allhutter et al., 2020]. However intuitive this may seem, it is nowhere substantively argued that statistical predictions of long-term unemployment from non-experimental data are reliable estimates for the effectiveness of labor-market 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.


Risk of Entry into and Probability of Exit from Unemployment
Running annual average December 2012 to December 2022 in Percent
Germany

Source: Statistik der Bundesagentur für Arbeit

Figure 2: Data and Figure from the German PES [Bundesagentur für Arbeit, 2023]. The risk of entering unemployment is estimated as the number of newly registered unemployed divided by the number of employees subject to social insurance contributions. The exit probability from unemployment is estimated as the number of registered unemployed who find a job in the primary labor market relative to the number of registered unemployed. Both time series are running annual averages from December 2012 to December 2022.

Indeed, 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 [Kleven et al., 2023, Bishu and Alkadry, 2016]. On the other hand, the gender gap in unemployment rates has largely disappeared over the last decades [Albanesi and Şahin, 2018]. Nevertheless, subtle 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 (see Figure 2). The obvious worry is that prediction algorithms will pick up on, entrench, or even exacerbate, these historical trends, as demonstrated in [Kern et al.,

[3] See Mueller and Spinnewijn [2023] for the economic perspective on predicting long-term unemployment.

[4] The focus on ALMPs restricts the set of policy options. ALMPs target *supply-side* problems by increasing human capital and *matching* problems by supporting job search. *Demand-side* policies might focus on the creation of jobs instead [Green, 2022].

2021]. The Austrian proposal for an LTU prediction algorithm furnishes a particularly dramatic example. That algorithm takes as input an explicitly gendered binary feature “obligation to care”, which has a negative effect on the predicted re-employment probability and, by design, can be set to 1 only for women [Allhutter et al., 2020]. 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. In the following section, we describe a methodology for predicting the evolution of reemployment probabilities from pre-deployment data.

3 Performativity and Prospective Fairness

First, some technicalities. 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, $Y^1, \dots, Y^{|\mathcal{D}|}$ 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:

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

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 [Levy et al., 2021]. 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. [2021].

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 [Perdomo et al., 2020]. 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:

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

Only the first term in this inequality is a quantity that can be directly estimated from training data. 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 will require more careful modelling. But careful modelling of the various ways in which predictions might influence decisions is precisely what we would like to encourage.

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:

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

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. Some of these effects are neither intended nor to be expected. For example, we assume that the conditional average treatment effects (CATEs) of the allocation on the outcome are stable across time:

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

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.

While *algorithmic* effects of deployment are intended and, to some degree, foreseeable, *feedback* effects are more complicated to model. Following Mishler and Dalmasso [2022] and Coston et al. [2020], we assume away the possibility of feedback effects, leaving these for future research:

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

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 $P_{\text{post}}(Y = y|A = 1) - P_{\text{post}}(Y = y|A = 0)$. Each of these are natural and important instances of $\varphi(\cdot)$. The proof is deferred to the supplementary material.

Theorem 1. *Suppose that* Consistency, Unconfoundedness, No Unprecedented Decisions, Stable CATE *and* No Feedback *hold. Suppose also that* $P_{post}(A = a) > 0$. *Then,* $P_{post}(Y = y|\ A = a)$ *is given by*

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

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

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 toy model. However, in realistic high-dimensional settings, the first term might be estimated by regression and the second by multivariate density estimation. Finally, $P_{\text{post}}(Y = y|A = a)$ could be estimated by integration of the plug-in estimates.

4 A Toy Model of a Public Employment Service

Our population of interest are the recently unemployed who have registered with some public employment service. For simplicity, we treat the gender variable A as binary. Obligation to care (X_1) is correlated with gender and increases the probability of long-term unemployment. In this model, the care-penalty is the only mechanism making gender a relevant axis of inequality. Educational attainment (X_2) is independent of gender and increases the probability of finding a job. Prior to the deployment of statistical profiling, the assignment into a labor market program is modelled as random, with 40% of the registered unemployed being allocated. This is consistent with empirical results by Lechner and Smith [2007], Goller et al. [2021], Cockx et al. [2023]. These variables determine Y_{Prior}, a binary variable that is 1 if the individual becomes unemployed in the long-term. High educational attainment, absence of care obligations, and participation in the labor market program all increase the reemployment probability.

$$
\begin{aligned}
A \in \{0,1\} &\sim \text{Bernoulli}(0.5), && 0 := \text{non-female;}\\
X_1 \in \{0,1\} &\sim \text{Bernoulli}(0.2 + 0.4A), && 0 := \text{no obligation to care;}\\
X_2 \in \{0,1\} &\sim \text{Bernoulli}(0.2), && 0 := \text{low educational attainment;}\\
D_{\text{Prior}} \in \{0,1\} &\sim \text{Bernoulli}(0.4), && 0 := \text{no ALMP, and}\\
Y_{\text{Prior}} \in \{0,1\} &\sim \text{Bernoulli}(0.5 + 0.3X_1 - 0.2X_2 - 0.2D_{\text{Prior}}) && 0 := \text{non-LTU.}
\end{aligned}
$$

Under the pre-deployment distribution, the gender reemployment gap is about 12 percentage points, with 56% of women and 44% of non-women becoming long-term unemployed. The overall population probability of becoming long-term unemployed is 50%.

Although its budget only allows the employment service to allocate 40% of the population to the program, it would like to make allocations more effective. To implement an algorithmic allocation policy, a logistic regression is trained on the features A, X_1, X_2 and the target variable Y_{Prior}. The resulting risk score R informs two potential policies, roughly resembling the Flemish policy of prioritizing the high-risk group and the Austrian policy of prioritizing the middle-risk group. Both policies fully automate allocation by thresholding the risk score. Under the Flemish-style policy, all and only individuals above the 60th risk percentile, t_F, are allocated into the program. Under the Austrian-style policy, the employment service restricts access to labor market programs to people above the 30th percentile $t_{\text{A-high}}$ and below the 70th percentile $t_{\text{A-low}}$. Due to sparse risk scores, the Austrian policy would allocate about 60% of the population into programs. To ensure that the share of treated stays constant at 40%, we multiply the resulting assignment by a Bernoulli random variable B parameterised by $0.4/0.6 = 2/3$. All the assumptions of the previous section are satisfied by design; the example respects the causal structure of Figure 1.

$$
\begin{aligned}
B \in \{0,1\} &\sim \text{Bernoulli}(2/3) && \\
D_{\text{A}} \in \{0,1\} &= \mathbb{1}[t_{\text{A-low}} \leq R \leq t_{\text{A-high}}] \times B && 0 := \text{no ALMP;}\\
Y_{\text{Post-A}} \in \{0,1\} &\sim \text{Bernoulli}(0.5 + 0.3X_1 - 0.2X_2 - 0.2D_{\text{A}}) && 0 := \text{non-LTU;}\\
D_{\text{F}} \in \{0,1\} &= \mathbb{1}[R \geq t_F] && 0 := \text{no ALMP, and}\\
Y_{\text{Post-F}} \in \{0,1\} &\sim \text{Bernoulli}(0.5 + 0.3X_1 - 0.2X_2 - 0.2D_{\text{F}}) && 0 := \text{non-LTU.}
\end{aligned}
$$

Neither the Flemish nor Austrian-style policies allocate anyone without care obligations and with high educational attainment $(X_1 = 0 \land X_2 = 1)$ to the program. Focusing on those at high risk, the Flemish-style policy assigns all and only those with care obligations to the program, whether female or not. Since 60% of women and 20% of non-women have care obligations, this policy treats precisely 40% of the population. Under the Austrian-style policy, women with low educational attainment but no care obligations $(X_1 = 0 \land X_2 = 0)$ and those with care obligations but high educational attainment $(X_1 = 1 \land X_2 = 1)$ receive a 66% chance of being allocated into the program;

it denies the program to all other women. All others, except those $(X_1 = 0 \wedge X_2 = 1)$, receive a 66% chance of being allocated into the program.

We would like to predict the overall reemployment probability, as well as the share of women and non-women that become long-term unemployed, after implementation. Analytically, we derive the following results for our toy model: the Flemish-style policy leaves the overall share of long-term unemployed unchanged (at 50%) while the Austrian-style policy slightly decreases long-term unemployment (to 49%). The Flemish-style policy brings the gender gap in long-term unemployment down from 12 to 4 percentage points by decreasing long-term unemployment among women $(P_{\text{post-F}}(Y = 1 \mid A = 1) = 52\%)$ and accepting an increased share (48%) among the rest. The gender gap increases under the Austrian policy to 17 percentage points. Under this algorithmic policy, women face higher long-term unemployment shares than before $(P_{\text{post-A}}(Y = 1 \mid A = 1) = 58\%)$, while the share among the others slightly decreases (41%). The detailed calculations are given in the Supplementary Material. Thus, it is possible to predict that (1) the Austrian-style policy will exacerbate the gender reemployment gap, (2) the Flemish-style policy will ameliorate it, and (3) neither will have a large effect on the population reemployment probability. Since both policies rely on the same predictive model R, these differences would not be visible to internal fairness metrics.

5 Conclusion and Future Work

The deployment of an algorithmically informed policy is an intervention into the causal structure of society that can have important performative effects. Therefore, we argue for a prospective evaluation of risk assessment instruments: comparing the relevant structural inequalities at training time with the situation likely to arise *after* the algorithmic policies are deployed. If the algorithmic policy changes decision making, it is likely to change the distribution of the outcome variable. That undermines static, group-based fairness notions that include the outcome variable. But even fairness notions that are not self-undermining in this sense give no answer to the fundamental question of fair machine learning: whether the deployment of an algorithmic policy will exacerbate structural inequalities.

In this paper, we develop such a prospective fairness methodology. We have shown that one can identify the effect of an algorithmic policy on a number of natural measures of structural inequality from the combination of (1) observational, pre-interventional data and (2) knowledge about the proposed policy. This result holds under a set of assumptions: UNCONFOUNDEDNESS of the potential outcomes with the policy decisions, STABLE CONDITIONAL TREATMENT EFFECTS, NO FEEDBACK, and NO UNPRECEDENTED DECISIONS. We illustrate the proposal with a toy model of a public employment service. Two potential policies, one of prioritisation and one of efficiency, are informed by predictions of the risk of long-term unemployment. We show that it is possible to predict that the former policy will ameliorate the gender reemployment gap, while the latter will exacerbate it.

Future research should extend this work and its limitations. On a theoretical level, it is important to consider weaker assumptions to allow for the analysis of more complex situations. Most importantly, methods from dynamical causal modelling can be used to relax the NO FEEDBACK assumption. Furthermore, axiomatic approaches to the measurement of inequality from the theory of social choice may help narrow down the set of admissible fairness metrics $\varphi(\cdot)$ and elucidate the trade-offs between them. Our toy model can be extended to situations in which (1) the pre-interventional assignment is not random but informed by caseworkers' decisions; (2) the algorithm and caseworkers use different inputs; (3) risk scores only inform, but do not fully determine, the allocation decisions; and (4) allocation into the programme has heterogeneous treatment effects. Future work could also utilise this model set-up for a systematic comparison of the effect of different static fairness constraints on structural inequalities. In the future, we would like to apply this methodology to real administrative data from public employment services.

Acknowledgments and Disclosure of Funding

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Sebastian Zezulka.

References

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

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

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

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

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

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

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

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

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

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

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

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

Lucius E. J. Bynum, Joshua R. Loftus, and Julia Stoyanovich. Counterfactuals for the Future, 2022.

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

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

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

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

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

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In Hal Daumé, III and Aarti Singh, editors, *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2185–2195. PMLR, 2020.

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

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

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

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

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

Daniel Goller, Tamara Harrer, Michael Lechner, and Joachim Wolff. Active labour market policies for the long-term unemployed: New evidence from causal machine learning, 2021.

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

Lily Hu and Yiling Chen. 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, 2018. doi: 10.1145/3178876.3186044.

Yaowei Hu and Lu Zhang. Achieving Long-Term Fairness in Sequential Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9549–9557, 2022.

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

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

Christoph Kern, Ruben L. Bach, Hannah Mautner, and Frauke Kreuter. Fairness in Algorithmic Profiling: A German Case Study, 2021.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Toru Kitagawa and Aleksey Tetenov. Equality-minded treatment choice. *Journal of Business & Economic Statistics*, 39(2):561–574, 2019. doi: 10.1080/07350015.2019.1688664.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores, 2016.

Henrik Kleven, Camille Landais, and Gabriel Leite-Mariante. The Child Penalty Atlas. *National Bureau of Economic Research*, 2023. doi: 10.3386/w31649.

Michael C. Knaus, Michael Lechner, and Anthony Strittmatter. Heterogeneous Employment Effects of Job Search Programs. *Journal of Human Resources*, 57(2):597–636, 2020. doi: 10.3368/jhr.57.2.0718-9615r1.

Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There?, 2021.

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

John Körtner and Ruben L. Bach. Inequality-Averse Outcome-Based Matching. 2023. doi: 10.31219/osf.io/yrn4d.

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

Michael Lechner and Jeffrey Smith. What is the value added by caseworkers? *Labour economics*, 14(2): 135–151, 2007.

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

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 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, 2019. doi: 10.24963/ijcai.2019/862.

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

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

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

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

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

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

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

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

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

Alexander Williams Tolbert and Emily Diana. Correcting Underrepresentation and Intersectional Bias for Fair Classification, 2023.

Davide Viviano and Jelena Bradic. Fair Policy Targeting. *Journal of the American Statistical Association*, pages 1–14, 2023. doi: 10.1080/01621459.2022.2142591.

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

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

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

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

Supplementary Material: Performativity and Prospective Fairness

Anonymous Author(s)
Affiliation
Address
email

A Proof of Theorem 1

Proof of Theorem 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: $P_{\text{post}}(Y = y \mid A = a) =$

Submitted to the Workshop "Algorithmic Fairness through the Lens of Time" at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

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

Next, note that, whenever defined,

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

Therefore,

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

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

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

as required. □

B Analytical Computations for Toy Model

Table 1: We tabulate the terms relevant for computing $P_{\text{postA}}(Y = 1|A = 1)$ and $P_{\text{postF}}(Y = 1|A = 1)$ according to Theorem 1. The eight column computes the products of terms in the third, fourth and fifth column. The ninth column computes the products of terms in the third, fourth and fifth column. Summing the terms in the eight column, we have that $P_{\text{postA}}(Y = 1|A = 1) = 58\%$. Summing the terms in the ninth column, we have that $P_{\text{postF}}(Y = 1|A = 1) = 52\%$.

Care	Ed.	D	$P_{\text{pre}}(Y = 1\|A, X, D = 1, x, d)$	$P_{\text{pre}}(X = x\|A = 1)$	$P_{\text{postA}}(D = d\|A, X = 1, x)$	$P_{\text{postF}}(D = d\|A, X = 1, x)$		
0	0	0	.5	.64	.334	1	.107	.32
0	0	1	.3	.64	.666	0	.128	0
0	1	0	.3	.16	1	1	.048	.048
0	1	1	.1	.16	0	0	0	0
1	0	0	.8	.16	.334	.043	.384	0
1	0	1	.6	.16	.666	.064	0	.096
1	1	0	.6	.04	.334	0	.008	0
1	1	1	.4	.04	.666	1	.011	.016

Table 2: We tabulate the terms relevant for computing $P_{\text{postA}}(Y = 1|A = 0)$ and $P_{\text{postF}}(Y = 1|A = 0)$ according to Theorem 1. The eight column computes the products of terms in the third, fourth and fifth column. The ninth column computes the products of terms in the third, fourth and fifth column.Summing the terms in the eight column, we have that $P_{\text{postA}}(Y = 1|A = 0) = 41\%$. Summing the terms in the ninth column, we have that $P_{\text{postF}}(Y = 1|A = 0) = 48\%$.

Care	Ed.	D	$P_{\text{pre}}(Y = 1\|A, X, D = 0, x, d)$	$P_{\text{pre}}(X = x\|A = 0)$	$P_{\text{postA}}(D = d\|A, X = 0, x)$	$P_{\text{postF}}(D = d\|A, X = 0, x)$		
0	0	0	.5	.32	.334	1	.053	.16
0	0	1	.3	.32	.666	0	.064	0
0	1	0	.3	.08	1	1	.024	.024
0	1	1	.1	.08	0	0	0	0
1	0	0	.8	.48	1	0	.384	0
1	0	1	.6	.48	0	1	0	.288
1	1	0	.6	.12	.334	0	.024	0
1	1	1	.4	.12	.666	1	.032	.048