A Proof of Proposition 3.9

Proof: By definition, the treatment $t^{(i)}$ is fair if and only if $t^{(i)} = s(\psi^{(i)})$.

"$\Leftarrow$", i.e. "$t^{(i)}$ is fair if $\hat{\varphi}(\tilde{\mathbf{x}}^{(i)}) = \psi^{(i)}$":

$$\begin{aligned} \hat{\varphi}(\tilde{\mathbf{x}}^{(i)}) &= \psi^{(i)} \\ \Rightarrow s(\hat{\varphi}(\tilde{\mathbf{x}}^{(i)})) &= s(\psi^{(i)}) \\ \Rightarrow t^{(i)} &= s(\psi^{(i)}) \end{aligned}$$

"$\Rightarrow$", i.e. "if $t^{(i)}$ is fair, then $\hat{\varphi}(\tilde{\mathbf{x}}^{(i)}) = \psi^{(i)}$":

$$\begin{aligned} t^{(i)} &= s(\psi^{(i)}) \\ \Rightarrow s(\hat{\varphi}(\tilde{\mathbf{x}}^{(i)})) &= s(\psi^{(i)}) \\ \overset{s() \text{ str. mon.}}{\Rightarrow} \hat{\varphi}(\tilde{\mathbf{x}}^{(i)}) &= \psi^{(i)} \end{aligned}$$ □

B Comparison to Other Concepts

In this section, we compare our proposed framework with two (seemingly) related notions that have been previously discussed in the literature: The legal concept of *substantive equality* in EU non-discrimination law together with the definition of *bias transforming fairness metrics* as introduced by Wachter et al. [74] and *counterfactual fairness* by Kusner et al. [48].

Relation to Substantive Equality and Bias Transforming Fairness Metrics. Wachter et al. [74] discuss the "legality of fairness metrics under EU non-discrimination law" and distinguish two normative concepts described in the EU non-discrimination law, namely *formal equality* and *substantive equality*. While in formal equality, the aim is "to not make society more unequal than the status quo", in substantive equality "true equality can only be achieved by accounting for historical inequalities which actively ought to be eroded". Building on this, they define concepts of *bias preserving fairness metrics* – that aim at formal equality – and *bias transforming fairness metrics* – that aim at substantive equality. Most interestingly, this can be related to our concept and the cases of no PAs (in Section 3.2) and with PAs (in Section 3.3), respectively: If no PAs are present, we aim at individual well-calibration which basically means that the ML model shall reproduce the true data-generating process as well as possible, i.e., perpetuating the status quo. However, if PAs are present, we account for historical inequalities – by moving to the FiND world – and actively erode them – by using warped world data for training and prediction. While our concept is essentially philosophically backed up by Wachter et al. [74], we contribute by translating this idea into an actionable framework for designing ML models.

Relation to Counterfactual Fairness. Counterfactual (Cf) fairness [48] also considers a fictitious world in comparison to the real world. While seemingly related at first glance, their concept differs substantially from our concept of FiND and warped worlds; see Figure 2 for a comparison with a very simple DAG, where the top left DAG shows the factual, real world. In our FiND world (top middle DAG), the PAs have no causal effect on the target and also not on features mediating this effect. The individual still belongs to its original class of the PA but with changed real-world descendants of the PA. In their Cf world (bottom left DAG), an individual belongs to a different class of the PA, which also affects descendants. This means the FiND world changes (some) features and

the target by modifying causal relationships regarding the PA, while the Cf world intervenes on the PA, maintaining real-world causal relationships. We then approximate the FiND world through the warped world, train an ML model, evaluate the model, and predict new observations with the model in the warped world (dashed box). Cf fairness trains an ML model in the factual, real world, and evaluates it by comparing its predictions using factual and counterfactual values for the PA and remaining features to assess fairness (dotted box).

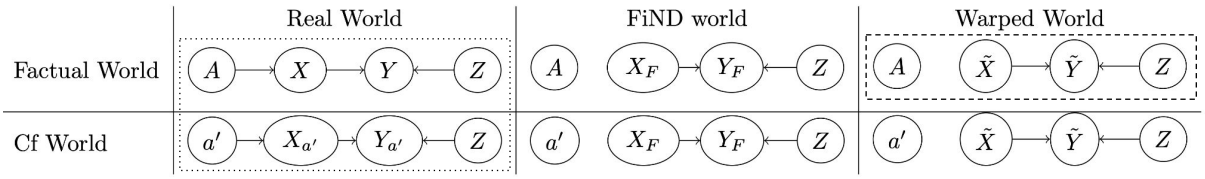


Fig. 2. Relating our concept to counterfactual (Cf) fairness by Kusner et al. [48] on DAGs comprising PAs (A), features descending from A (X), other features (Z), and a target Y. We argue to train, evaluate, and predict in the warped world (dashed box), while Cf fairness compares predictions in the factual world and the Cf world (dotted box).

Despite these differences, we fulfill Cf fairness in a certain sense: Consider a Non-White person i. Cf fairness demands that its prediction (using its factual feature vector $\mathbf{x}^{(i)}$) is the same as for its counterfactual j, where the person belongs to class White, i.e., using $\mathbf{x}^{(j)} = \mathbf{x}^{(i)}_{a'}$ ($\neq \mathbf{x}^{(i)}$ in general). Since these two fictitious persons have the same task-specific merit and hence the same FiND-world feature vectors $\mathbf{x}^{(i)}_F = \mathbf{x}^{(j)}_F$ in our concept, we demand that they have the same warped world feature vectors $\tilde{\mathbf{x}}^{(i)} = \tilde{\mathbf{x}}^{(j)}$, which leads to equal predictions $\hat{\varphi}(\tilde{\mathbf{x}}^{(i)}) = \hat{\varphi}(\tilde{\mathbf{x}}^{(j)})$. In this world, the PA has no effect, which means that $\tilde{\mathbf{x}}$ does not contain descendants of the PA. With Lemma 1 of Kusner et al. [48], this fulfills Cf fairness. However, as shown above, this is just a necessary condition for fairness and a form of individual well-calibration is paramount for achieving fairness. Measuring predictive performance must take place in the "correct" world, i.e., in the real world (if no PAs are present), or in the FiND world – approximated by the warped world – (if PAs are present).

C Implementation of the Concept

In this section, we illustrate how our framework could be used for tackling fairness issues in an applied use case. Section C.1 outlines how to assess the FiND world via warping and how to train and predict in the warped world. Section C.2 presents illustrative experiments with the COMPAS data. In the illustrative experiments below, we assume that the DAG is correct and that the warping is valid. We then focus on descriptive evaluations, comparing real-world and warped-world predictions for individuals, identifying the most discriminated individuals, and comparing these values on the level of PA-groups.

C.1 Warping via Rank-Preserving Interventional Distributions

We have observed so far that for valid training of a model as well as for its evaluation regarding fairness, a warping approach must be found to create an approximation of the FiND world. This can be considered to be a pre-processing approach to fairML [for more details on categorizing fairML approaches, see 15]: First, a warping function is learned by a suitable warping method and applied to the entire data. Then, a predictive model is tuned,

trained, and evaluated on the warped data. Finally, at prediction time, the feature vector of a new observation is warped by the same warping function and the predictive model is applied to that warped feature vector.

In the following, we use rank-preserving interventional distributions (RPID) to identify the FiND world and a residual-based warping method, both proposed by [12] as warping method. We sketch the idea in the following and refer to the original source for details. This is just one example of a possible warping method [another option could be to adapt quantile regression forests as proposed by 64], and it is left to future research to develop and compare more warping methods which can be embedded into our framework.

The core idea of RPID is to make the variables neutral with respect to the PA while maintaining any individual merits, sequentially, for all descendants of the PA. For example, a Non-White person, who is at the 90%-quantile of the confounder-specific distribution of X_P in the Non-White subgroup will be "warped" to the 90%-quantile of the confounder-specific distribution of X_P in the White subgroup, see also Figure 3.

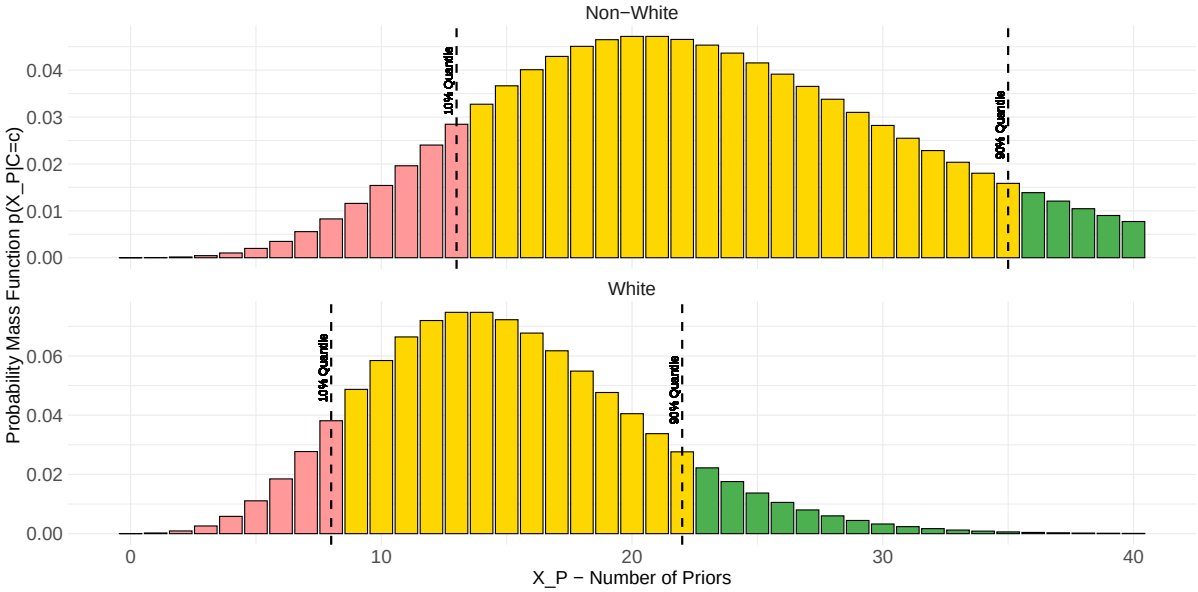


Fig. 3. Illustration of RPID: Quantiles (here: 10% and 90%) of the protected group (here: Non-White, top plot), are warped to their counterparts in the non-protected group (here: White, bottom plot). Fictitious distributions for illustration purposes.

RPID defines interventions on the SCM that lead to a FiND world that respects both the requirements on the PAs and individual merits. The estimands are defined on the joint counterfactual post-intervention distribution, e.g., the counterfactual X_P-distribution. Whether the desired estimand can be identified, i.e., expressed as a function of the observed data, requires considerations on, e.g., confounders of the mediator-outcome relationship. The concrete estimation method is based on a particular g-formula factorization to estimate the distributions of the FiND world. A model-based residual approach is used for estimating FiND world distributions and for warping real-world data: Comparably to [12], we assume an additive, homoscedastic error term in all models of mediators and target variable, given parents of these variables in the DAG, where we consider separate models per subgroup of the PA.

This ultimately leads to a warping function from the real world to the warped world. Model training, validation, and prediction for new observations take place in the warped world. This means that we focus on estimating a well-performing model, and try to achieve individual well-calibration – but this time in the warped world. There is no longer a need for special evaluation measures for predictive performance in order to assess (un)fairness. The complexity is shifted to estimating the SCMs and the individual conditional quantiles. Note that it is of paramount importance for a successful practical use case to implement a close collaboration between subject-matter experts and machine learning experts, since expert knowledge will have to go hand-in-hand with empirical findings [as also pointed out by others before, e.g., 26, 39, 46, 53] in order to derive realistic causal models. The practical usability of this approach depends on the extent to which the causal relationships in the real world can be estimated as well as on the success of removing certain causal relationships without modifying others.

C.2 Experiments

As an example, we show experiments on the COMPAS data [2].[4] These experiments illustrate how our framework could be applied in practice and show how the warping approach affects the predictions for the different subgroups regarding the PA. However, this must not be misunderstood as a fully applied use case. As Section D in the Appendix demonstrates, several normative questions must be answered carefully, followed by cautious modeling of the data. We use the DAGs of Figure 1b, noting again that this is just an illustration, and the true DAG will be more complex, contain more features, and must be developed together with domain experts.

We apply the residual-based warping approach outlined above: We sequentially warp X_P, X_D, and Y of the Non-White individuals, using Poisson models for $X_P|A=a,C$, and logit models for $X_D|A=a,C$ and $Y|A=a,C,X_D,X_P$, with $a \in \{\textit{Non-White}, \textit{White}\}$ (see Figure 1b for variable descriptions). In a nutshell, this works as follows, at the example of X_P: We train two Poisson models $\hat{f}_n(C)$ and $\hat{f}_w(C)$, for the Non-White and White subgroups, respectively. Then, we warp Non-White individual's i value $x_P^{(i)}$ in three steps: (i) compute its residual with respect to the Non-White model, i.e., $r^{(i)} = x_P^{(i)} - \hat{f}_n(c^{(i)})$; (ii) compute its relative position among all Non-White residuals for this model, e.g., 95% quantile, and find the corresponding quantile $q_w^{(i)}$ of White residuals for the White model $\hat{f}_w(\cdot)$; (iii) warp $x_P^{(i)}$ to the sum of this transformed residual and the White model prediction, i.e., $\tilde{x}_P^{(i)} = \hat{f}_w(c^{(i)}) + q_w^{(i)}$. This is done sequentially for X_D and Y as well, where the already warped values for X_D and X_P ($\tilde{x}_D^{(i)}$ and $\tilde{x}_P^{(i)}$) are plugged into the prediction model of Y in step (iii).[5] As result, i's features are warped from $\mathbf{x}^{(i)}$ to $\tilde{\mathbf{x}}^{(i)}$ and its target from $y^{(i)}$ to $\tilde{y}^{(i)}$.[6]

White individuals' features and target are not warped, i.e., $\tilde{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)}$ and $\tilde{y}^{(i)} = y^{(i)}$, but the predicted recidivism rates also change from real-world rates $\hat{\pi}^{(i)}$ to warped-world rates $\hat{\varphi}^{(i)}$ due to differing training data of Non-White individuals. As prediction models for the recidivism rate, we train logit models on real-world and warped data, respectively. We train all models (warping and recidivism rates) on an 80% random subsample of the data and test on the remaining observations. Table 2 shows the individuals of the test data where the predicted recidivism rate changed most between real world and warped world (column *diff*). We can observe that (a) this is associated with a decrease from X_P to $\tilde{X}_P$ for Non-White individuals who had the strongest change, and also that (b) the predicted

[4] All experiments are fully reproducible, GitHub repository is revealed upon acceptance.

[5] Technical note: While we use the exact warped features in this sequential algorithm, we show rounded versions in Table 2 for better readability.

[6] At prediction time, where target $y^{(i)}$ is not available, we only warp the feature vector $\mathbf{x}^{(i)}$.

Table 2. Strongest warped individuals for (a) Non-White and (b) White individuals in test data, where $diff = \hat{\varphi}^{(i)} - \hat{\pi}^{(i)}$. Individuals above the dots are negatively discriminated in the real world, while individuals below the dots are positively discriminated in the real world. Overall, negative discrimination is higher for Non-White individuals and positive discrimination is higher for White individuals.

(a)

gender	*age*	X_P	X_D	$\tilde{X}_P$	$\tilde{X}_D$	$\hat{\pi}^{(i)}$	$\hat{\varphi}^{(i)}$	*diff*
Male	63	23	0	15	0	0.83	0.30	-0.53
Male	53	26	0	18	0	0.93	0.41	-0.52
Male	56	22	0	15	0	0.85	0.35	-0.50
...	...	...	...	...	...	...	...	...
Male	54	0	1	0	1	0.15	0.22	0.07
Male	55	0	1	0	1	0.15	0.22	0.07
Male	56	0	1	0	1	0.14	0.21	0.07

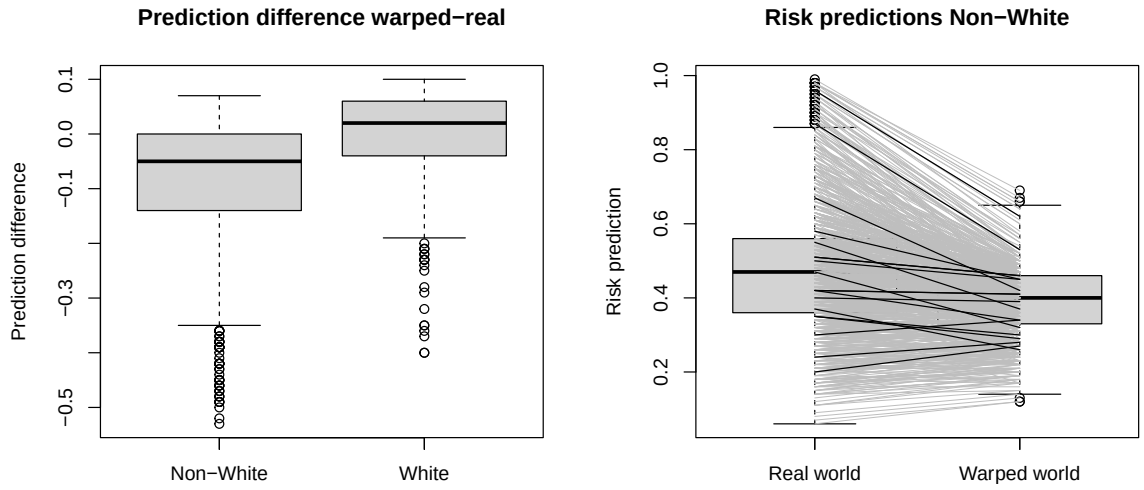
(b)

gender	*age*	X_P	X_D	$\hat{\pi}^{(i)}$	$\hat{\varphi}^{(i)}$	*diff*
Female	50	30	1	0.94	0.53	-0.41
Female	47	28	0	0.94	0.54	-0.40
Male	55	33	0	0.97	0.60	-0.37
...	...	...	...	...	...	...
Male	52	0	1	0.17	0.26	0.09
Male	52	0	1	0.17	0.26	0.09
Male	53	0	1	0.16	0.26	0.10

recidivism probability is lower in the warped world for White individuals with a high value of X_P (values range between 0 and 38 in the data). This is due to a lower coefficient in the logit model related to this feature, which is 0.15 (se 0.0076) in the real world and 0.05 (se 0.0075) in the warped world.

Figure 4a visualizes the prediction differences $\hat{\varphi}^{(i)} - \hat{\pi}^{(i)}$ between the worlds for both PA groups. While White prediction differences vary around 0 (mean: -0.0047, p-value of t-test against 0: 0.254), Non-White prediction differences show a trend towards lower values (mean: -0.0918, p-value of t-test: $< 10^{-91}$), which means that – on average – their predicted recidivism rates are 9.2% points lower in the warped world. While 76% of Non-White individuals receive lower predictions in the warped world (41% for White), 24% receive higher predictions (59% for White). Figure 4b visualizes the predictions in both worlds for the Non-White subgroup. We see the overall negative trend, where the ranking within the group changes between the two worlds. This indicates differing individual strengths of discrimination in the real world and illustrates the individual perspective of our proposal.

To illustrate the impact that the concrete choice of warping method can have on the results, we also applied "fairadapt" [64] via the R package `fairadapt` [63]: The goal is to compute *fair twins* via a *do*-intervention on the PA, aiming to minimize "the distortion in the data coming from the projection", using quantile regression

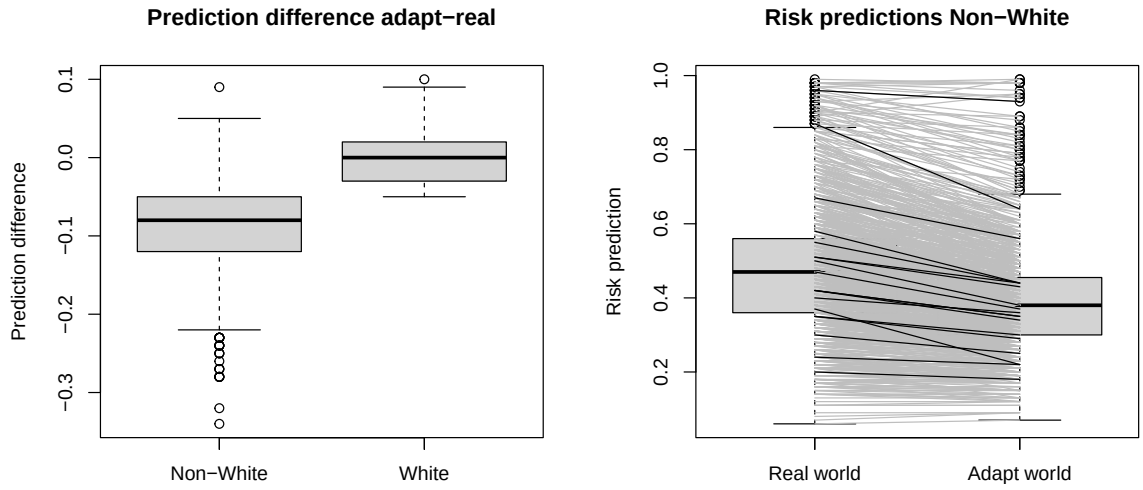


(a) (b)

Fig. 4. Recidivism predictions in the warped world and the real world. (a) shows differences in predictions for both groups and (b) shows predictions in both worlds for the Non-White group, where lines connect the same individuals (a random subsample has black lines for better readability).

forests [56] as estimation method. Figure 5a shows the prediction differences between the real world and the "adapt" world for both PA groups. Similar to above, the prediction difference for the Non-White subgroups shifted toward negative values, with the values for the White subgroup varying around 0. Figure 5b shows the predictions in the real world and the adapt world for the Non-White subgroup. We see an overall negative trend as above, where the variability seems to be higher in the adapt world than in the warped world.

In an applied use case, we cannot prove or disprove that the warping worked and that the warped world is a good approximation of the FiND world. Similarly, we cannot judge whether the residual-based warping approach or fairadapt is superior. To make such a statement, we would need access to the FiND world. To properly investigate whether a warping method works in this sense, a thorough simulation study is necessary. Such a simulation study should also consider investigating other warping methods and comparing their behavior. Answering the question of how to optimally design a warping method to approximate the FiND world is important future work and beyond the scope of this paper. Nevertheless, the experiments presented here illustrate the type of results that can be expected from applying our framework in practice. Furthermore, the fact that the general direction of the results of the two warping approaches analyzed here are similar is an indication that both capture real-world discrimination and can potentially serve as bias-transforming methods.



(a) (b)

Fig. 5. Pendant to Figure 4, showing recidivism predictions in the adapt world and the real world. (a) shows differences in predictions for both groups and (b) shows predictions in both worlds for the Non-White group, where lines connect the same individuals (a random subsample has black lines for better readability).

D Workflow for Applied Use Case

Figure 6 shows a workflow of how to use our framework in an applied ADM use case. Three normative questions arise, although it is important to note that these can never be answered empirically:

(1) How is the task-specific merit $m^{(i)}$, i.e., the measure of task-specific equality of individuals, defined?
(2) Which attributes are defined as PAs (if there are any)?
(3) Shall the treatment function $s(m^{(i)})$ be modified, and if so, how?

(1) Task-specific merit: For the task at hand, the normative decision must be made of which features m are used to define individuals as equal. In other words, two individuals i and j are considered equal *for this specific task* if and only if $m^{(i)} = m^{(j)}$ and, hence, have the same task-specific merit. The task-specific merit may consist of (a function of) observable $v^{(i)}$ and unobservable $z^{(i)}$ features, $m^{(i)} = f(v^{(i)}, z^{(i)})$. In the COMPAS example, $v^{(i)}$ might be the type of crime and $z^{(i)}$ might be the two-year recidivism rate. In an example where individuals apply for credit in a bank, $v^{(i)}$ might be the amount of credit and $z^{(i)}$ the probability of default. In a tax example, there might be no unobservable variables because all features for the tax declaration are observable at decision time.

(2) Protected attributes: Next, the normative decision must be made if there are PAs – and if there are any PAs, to which feature they correspond. If applicable, the decision can be made to just remove some path-specific effects starting in the PA and to keep others in the FiND world (aiming for lawful indirect discrimination). Together,

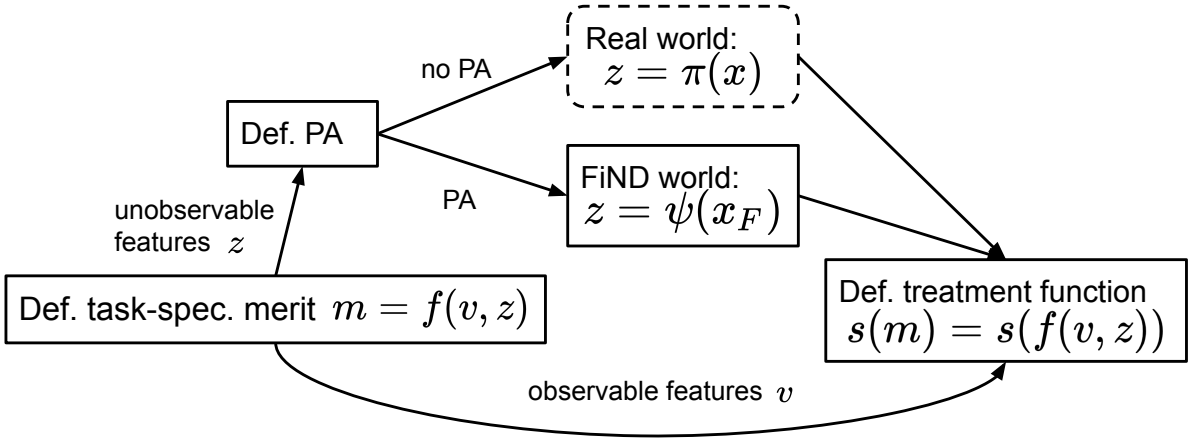


Fig. 6. Workflow for an applied use case. Solid rectangles refer to normative decisions.

this defines the FiND world. These are delicate questions and surely the result of a broader, societal discussion; the presence of PAs means that for some groups, the unobservable features are not used as-is, but they will be transformed into their FiND world counterpart. This might potentially change the (numerical) task-specific merit of some or all individuals and have a large impact on how people are ranked overall (as exemplified in the experiments in Section C.2). As a result, without PAs, the unobservable features will be identified with the estimand in the real world, i.e., $z^{(i)} = \pi(\mathbf{x}^{(i)})$, whereas in the presence of PAs, the unobservable features will be identified with the estimand in the FiND world, i.e., $z^{(i)} = \psi(\mathbf{x}_F^{(i)})$.

(3) Treatment function: Finally, the treatment function $s(m^{(i)}) = s(f(v^{(i)}, z^{(i)}))$ must be defined normatively. Again, this results from a societal or expert discussion regarding the use case at hand. In the COMPAS example, a binary decision on concrete treatment might be made using a threshold on the recidivism probability, where the threshold varies between different types of crime. In the credit example, a binary decision on granting credit might be made using a threshold on the probability of default, where the threshold varies between different amounts of credit. In the tax rate example, the decision only relies on observable features and is subject to regular political discussion.

Once these questions are answered, algorithms as outlined in Section C can be used to approximate the FiND world via warping. Then, warped-world data are used to train ML models and to predict from these trained ML models for new, unseen individuals – after warping their feature vector.

In summary, the task of subject matter researchers is to define the PAs, the DAG in the real world, and other variables that are part of the task-specific merit (outside of the DAG), where the tasks of ML experts are to find good warping methods to approximate the such defined FiND world and to train well-performing prediction models in the warped world.

Remark on PAs and observable features. Figure 6 shows that observable features are used as-is, no matter if there are PAs. However, the normative result of a discussion on this could be that not only the unobservable but also the observable features may be subject to historic unfairness (e.g., income with the PA gender) and that these features should also be transferred to the FiND world (e.g., asking "what would the income be in a world where gender has no causal effect on income"). This is a straightforward extension of our framework: Instead of using the observable feature v directly, it will be replaced by its FiND world counterpart v_F, which then must be estimated, e.g., by an ML model, yielding $\hat{v}$.

E Ethical Statements

Ethical Considerations Statement. Our work has a rather theoretical, philosophical perspective, and we do not propose a concrete application that could harm individuals. One goal of this paper is to disentangle normative choices and ML-related methodology to provide greater linguistic clarity when discussing fairness in ML and ADM systems. While conducting this work, it was important for us to be as specific as possible about this distinction between normative and methodological questions. We emphasize that the answers to these normative questions (as summarized in Section 4) must be the product of careful ethical considerations and, in many cases, the result of a broader societal discussion.

Researcher Positionality Statement. Part of the author group is educated and based in the fields of statistics and computer science. While working in methodological research of ML, the authors appreciate philosophical questions surrounding the analysis of data and are skeptical of technical proposals that seem not to deeply interrogate *why* they are proposing these solutions. The other part of the author group is educated and based in the fields of law and philosophy. They work on the very philosophical foundations of law and, thereby, also engage closely with ethical questions. The interdisciplinary nature of our team contributes to a more comprehensive understanding of fairness in ML.

Adverse Impact Statement. As for other technical solutions, blindly trusting techniques to overcome fairness problems can have negative societal impacts. However, we hope that by explicitly tasking non-ML users with answering three broadly understandable normative questions and by linking the ML model evaluation to the respective answers, we contribute a step in the direction of bringing societal needs and technical solutions closer together. Nevertheless, if the answers establish undesirable, e.g., discriminatory norms, then there is no technical possibility to remedy this. Therefore, it is of paramount importance that the normative stipulations – especially in critical applications – are transparent and result in a broad societal discussion.