Counterfactual Situation Testing: Fairness given the Difference

Jose M. Alvarez
University of Pisa & Scuola Normale Superiore
Pisa, Italy
jose.alvarez@sns.it

Salvatore Ruggieri
University of Pisa
Pisa, Italy
salvatore.ruggieri@unipi.it

Abstract

We present counterfactual situation testing (cfST), a new tool for detecting individual discrimination in datasets that operationalizes the Kohler-Hausmann Critique (KHC) of "fairness given the difference". In standard situation testing (ST), as with other discrimination analysis tools, the discrimination claim is recreated and thus tested by finding similar individual profiles to the one making the claim, the complainant c, and constructing a control group (*what is*) and a test group (*what would have been if*) of protected and non-protected individuals, respectively. ST builds both groups around c, performing an idealized comparison. Implicitly, from an individual fairness perspective, it assumes that it is fair to compare observably similar protected and non-protected individual profiles to test for discrimination. This approach is wrong under KHC as it fails to account for the pervasive effects of the protected attribute on the relevant attributes used for the decision in question. Under cfST, we extend ST by constructing the control group around the complainant, which is the factual, and the test group around its counterfactual using the abduction, action, and prediction steps for a given structural causal model. Unlike ST, we do not assume the same within-group ordering across protected and non-protected groups. We thus compare groups of not so similar individual profiles: one based on what we observe about c versus one based on a hypothetical, counterfactual representation of c. We compare cfST to existing ST methods along with counterfactual fairness using synthetic data for a loan application process. The results show that cfST detects a higher number of individual discrimination cases than the other methods.

1 Introduction

Discrimination, at least at an individual level, is conceived often as a causal claim on the effect of gender, race, or any other protected attribute on the decision outcome [21]. Most tools for detecting discrimination, including *situation testing* (ST), implement with varying degrees a conceptual experiment in which the protected attribute varies but all else remains the same [39, 6]. In ST, we replicate the decision process in question by passing through it similar individual profiles to the one making the discrimination claim (*the complainant*), and compare the outcomes of those that, like the complainant, belong to the protected group (the control group) against those that do not (the test group). ST is used today as a legal tool both in the USA [5] and the European Union [40].

Defining what constitutes "similar individual profiles" is the main challenge for ST and other tools as, conceptually, only by comparing (almost) identical units we are then able to detect the "treatment effect" of the protected attribute on the individual in question. In algorithmic ST [45, 49] the focus is on defining a distance function to measure similarity between the complainant and the corresponding control and test groups. This push for formalizing similarity has led to recent critiques against

Preprint. Under review.

the underlying causal model of discrimination (e.g., [23, 26]). [28], in our view, inspired and still best encapsulates these critiques. The underlying causal model is wrong because it reduces the protected attribute to a phenotype, ignoring the historical processes such attribute represents and how systematically pervasive it can be on all other attributes. This critique, which we refer to as the *Kohler-Hausmann Critique* (KHC), can be summarized as "fairness given the difference" between individuals.[1]

In this paper, we present a new algorithmic tool for discrimination analysis based on individual instances in a dataset of interest, *counterfactual situation testing* (cfST), by revisiting ST under the KHC. The current ST problem constructs and centers both control and test groups around the complainant using the same distance function. This approach is wrong based on the KHC since, knowing what we know about the systematic effects of the protected attribute on everything else, we cannot expect the control and test groups to be derived in a similar way. Under cfST, we instead construct the control group around the complainant, which is by definition the factual, and the test group around the complainant's counterfactual using the same distance function. This is because, under the dominant causal model of discrimination [28], the control group answers to the factual question of *what is* while the test group answers to the counterfactual question of *what would have been if*. We generate the counterfactuals using the abduction, action, and prediction steps [35] based on a structural causal model that describes the data generating model behind the dataset in question.

Our goal with cfST is to provide a meaningful discrimination tool that can be used for auditing (human or algorithmic) decision-making systems while addressing the fairness concerns around the dominant discrimination causal model. Using a synthetic dataset, we test our cfST tool and compare it to other ST and causal fairness methods. Experimental results show that cfST detects a higher number of individual discrimination cases, highlighting the impact of changing to a new conception of detecting discrimination based on fairness given the difference rather than idealized comparisons of individuals.

A motivating example. To illustrate the interplay between ST, KHC, and cfST, we present a scenario based on [32]. Suppose Clara, who was denied tenure due to having only 12 published papers, files a discrimination claim against her university. We then must consider the relevant information used for the decision in question (number of publications) along with other available information linked to the protected group (being female) in this context, such as Clara being a mother. Under ST, suppose we match Clara with Mike, who also published 12 papers and is a parent. If Mike did get tenure, would it disprove Clara's claim? Overall, would Mike be a fair counterfactual representation of Clara? Under KHC, no because we are aware that parenthood affects male and female researchers differently and we are thus not being fair given the known systemic differences between these two groups. Our view, however, might change if we added the information that Mike is a single parent. Why? Because then Mike would have experienced gender through parenthood in a way that is closer to how Clara experienced it as a female researcher: despite being male, Mike's parenthood would have hindered his research output in a female-like way. Hence, we would want Clara's counterfactual profile to reflect the lack of negative effects parenthood can have on male researchers' number of publications. Under cfST, assuming a measurable penalty for being a mother, we would expect to find a different kind of similar male profile to Clara, a counterfactual like, suppose, Vincent, who has 20 instead of 12 published papers.

Related work. *Discrimination discovery* [36] operationalizes traditional discrimination analysis tools [39, 6], such as audit studies [15] and correspondence studies [7], using algorithmic methods. Two approaches within this field use the legal ST framework [5, 40]: [45] use k-nearest neighbor to derive the control and test groups, while [49] use the weights of a structural causal model. More recent approaches that resemble ST, such as the FlipTest by [8] that uses optimal transport and [38] that use propensity score weighting, have appeared. Unlike cfST, these approaches center the construction of both the control and test groups around the factual instance. What separates cfST from these tools is our explicit attempt to address the alleged wrongness of the dominant causal model of discrimination [28] by constructing the test group on the counterfactual instance.

Causal fairness methods use structural causal models [33] to formalize, measure. and mitigate algorithmic bias (see [30] for a recent survey). Our proposed cfST tool falls within those causal

[1] The phrase does not appear in said paper, though it was mentioned by the author during a panel discussion at AFCR2021, which best captures her overall critique toward the causal model of discrimination.

fairness works that use (structural) counterfactuals as conceived by [35]. Among these works, cfST relates to counterfactual fairness [29] as we draw the counterfactuals for each individual given a decision-making model and compare it the corresponding factuals. However, we go a step further than CF by comparing similar profiles to the factual (the control group) and the counterfactual (the test group) of an individual instances to align more to how discrimination is tested in court. As we show later with out experiment, it is possible to be counterfactually fair based on the literal comparison of the factual and counterfactual instances, while still test positive for discrimination.

Individual fairness [14] states that that similar individuals should be treated similarly. The notion of similarity, given some distance function to define it, is central to ST and discrimination analysis. The KHC and, thus, cfST challenge this notion under "fairness given the difference" as we want "dissimilar" individuals to be treated similarly for a specific decision context.

2 Background

Let $\mathcal{D}$ denote the dataset of interest that contains the set of j non-protected, relevant attributes $\mathbf{X} = \{X_1, \dots, X_j\}$, the protected attribute A, and the predicted decision outcome $\widehat{Y}$. Let b represent the human or algorithmic decision-maker behind $\mathcal{D}$. We describe $\mathcal{D}$ as a collection of n tuples $\mathcal{D} = \{(x_{l,1}, \dots, x_{l,j}, a_l, \widehat{y}_l)\}_{l=1}^{n}$, with each l^{th} tuple $(\mathbf{x}_l, a_l, \widehat{y}_l)$ representing an individual profile. We assume A and $\widehat{Y}$ to be binary, such that $A = 1$ represents membership to the protected group while $\widehat{Y} = 1$ a positive decision outcome. Although here we focus on a single protected attribute, we can handle a set of d protected attributes $\mathbf{A} = \{A_1, \dots, A_d\}$.

Types of discrimination. We focus on the EU legal context, in which discrimination is classified as either direct or indirect. Direct discrimination occurs when the individual is treated less favorably on grounds of membership to a protected group, while indirect discrimination occurs when an apparently neutral practice disadvantages individuals that belong to a protected group. We are interested in *indirect algorithmic discrimination* for two reasons. First, unlike US law (disparate impact) [4], the decision-maker can still be liable for it despite lack of premeditation. All practices need to consider potential indirect discrimination implications.

Second, most EU legal scholars consider indirect discrimination the potential most common form of algorithmic discrimination as regulators are reluctant to allow A as a model input, dismissing direct discrimination. See [20] for details. Also, see [1] for a recent legal counterargument for this approach toward algorithmic discrimination. Our focus on indirect discrimination implies that the decision is based on non-protected attributes, namely $\widehat{Y} = b(\mathbf{X})$, though it does not rule out awareness about $\mathbf{A}$ as this is needed for detecting discrimination claims [27].

Structural causal models. We treat all recorded attributes in $\mathcal{D}$ as random variables, where for the j attributes we have a joint distribution $\mathbb{P}(\mathbf{X})$. A *structural causal model* (SCM) [33] $\mathcal{M} = \{\mathcal{S}, \mathcal{P}_{\mathbf{U}}\}$ allows us to factorize $\mathbb{P}(\mathbf{X})$ by describing how the set of endogenous (observed) variables $\mathbf{X} = \{X_1, \dots, X_j\}$ is determined from a set of exogenous (unobserved) variables $\mathbf{U} = \{U_1, \dots, U_j\}$ with prior distribution $\mathcal{P}_{\mathbf{U}}$ via the set of *structural equations* $\mathcal{S}$:

$$\mathcal{M} = (\mathcal{S}, \mathcal{P}_{\mathbf{U}}), \quad \mathcal{S} = \{X_i := f_i(X_{pa(i)}, U_i)\}_{i=1}^{j}, \quad \mathcal{P}_{\mathbf{U}} = \mathbb{P}(U_1) \times \cdots \times \mathbb{P}(U_j) \tag{1}$$

where each X_i is assigned a value through a deterministic function f_i of its causal parents $X_{pa(i)} \subset \mathbf{X} \setminus X_i$ and its corresponding noise variable U_i with distribution $\mathbb{P}(U_i)$. We consider the associated *causal graph* $\mathcal{G} := (\mathcal{V}, \mathcal{E})$, where each node $V_i \in \mathcal{V}$ represents an observed X_i variable and each directed edge $E_{(i,j)} \in \mathcal{E}$ represents the child-parent or cause-effect relation.

We make four assumptions common within the causal fairness literature [30]. First, we assume *causal sufficiency*, meaning there are no hidden confounders in $\mathcal{M}$. Second, we assume $\mathcal{G}$ to be *acyclical*, which makes $\mathcal{G}$ into a directed acyclical graph (DAG). Third, we assume *additive noise models* (ANM), where the structural equations in $\mathcal{S}$ are of the form:

$$\mathcal{S} = \{X_i := f_i(X_{pa(i)}) + U_i\}_{i=1}^{j} \tag{2}$$

which insures an invertible class of SCM (e.g., [22]). Fourth, we assume a *linear Gaussian ANM* as a modeling starting point that can change depending on what we are willing to assume about $\mathcal{D}$.

Generating counterfactuals. Under cfST, for a given SCM $\mathcal{M}$, we want to run *counterfactual queries* (third run on the causation ladder [34]) in which we answer *what would have been if* questions around the protected attribute A. Here, we want to evaluate the individual-level effect on (the factual) $\mathbf{X}$ after setting (the factual) A to the non-protected group using the *do-operator* [33]: $do(A := \alpha)$. Let $\mathbf{X}^{CF}$ denote the set of (structural) counterfactual variables we wish to generate from such intervention.[2] We generate the *counterfactual distribution* $\mathbb{P}(\mathbf{X}^{CF}_{A \leftarrow \alpha}(\mathbf{u}) \mid \mathbf{X}^F, A^F)$ via three steps [35]. *Abduction*: for each prior distribution $\mathbb{P}(U_i)$ that describes U_i, we compute the corresponding posterior distribution given the evidence, $\mathbb{P}(U_i \mid \mathbf{X}^F, A^F)$. *Action*: we intervene A using the set of interventions $\mathcal{I} = \{\alpha\}$ via $do(A = \alpha)$, obtaining a new SCM $\mathcal{M}^{\mathcal{I}}$. *Prediction*: we introduce the values from $\mathbb{P}(U_i \mid \mathbf{X}^F)$ into $\mathcal{M}^{\mathcal{I}}$ and compute the implied distribution.

Given (1) and (2), we follow the procedure for generating counterfactuals in Karimi et al. [25]. In particular, for the abduction step, we solve for U_i using the evidence $\mathbf{X}_i$ via $U_i = X_i - f_i(X_{pa(i)})$. This step is an individual-level statement on the residual variation under SCM $\mathcal{M}$, meaning it accounts for all that our assignment functions f_i, which are at the population level, cannot explain.

3 Counterfactual situation testing

The goal of *counterfactual situation testing* (cfST) is to construct and compare for each complainant c in the dataset $\mathcal{D}$ a control and a test group that embody the Kohler-Hausmann Critique (Section 3.1). Let $(\mathbf{x}_c, \mathbf{a}_c, \widehat{y}_c)$ represent the individual profile or *tuple of interest* under the discrimination claim. Individuals in the control group belong to the protected group while individuals in the test group belong to the non-protected group. Only individuals that belong to the protected group, $A = 1$, are eligible to be a complainant. We thus partition the dataset $\mathcal{D}$ under analysis into *protected*, $\mathcal{D}_1 \equiv \{l \in \mathcal{D} : a_l = 1\}$, and *non-protected*, $\mathcal{D}_0 \equiv \{l \in \mathcal{D} : a_l = 0\}$, *search spaces*. It follows that $c \in \mathcal{D}_1$.

We proceed as follows in cfST. First, we generate the counterfactual dataset $\mathcal{D}^{CF}$, via $do(A := 0)$, from $\mathcal{D}$. Now for each complainant c in $\mathcal{D}$, we use the factual tuple of interest to search over $\mathcal{D}_1$, while we use its corresponding counterfactual tuple to search over $\mathcal{D}_0$ (Section 3.2). Second, we use the distance function (Section 3.3) and the k-nearest neighbors algorithm (Section 3.4) to construct the k-sized control and test groups for c. Third, we statistically compare both groups to test for the individual discrimination claim of c (Section 3.5). See Algorithm 1 in Appendix A for an overall view of the steps within cfST.

3.1 Fairness given the difference: the Kohler-Hausmann Critique

Kohler-Hausmann [28] argues that the dominant model for proving discrimination, the "counterfactual causal model", is wrong because it reduces the protected attribute A into a phenotype in order to create a conceptual experiment in which individual units can be considered similar while adjusting for A to capture its "treatment effect". Her critique goes beyond the standard manipulation concern in the social sciences [3], in which A is an attribute that should not be modified. Instead, she subscribes to a *constructivist view* of A. Under this view, A simultaneously serves as an attribute and interpreter that fundamentally structures all other attributes of an individual. When A changes, other attributes in $\mathbf{X}$ most likely change as well. Therefore, under this view, it is unfair to construct a control group, which answers to *what is*, and a test group, which answers to *what would have been if*, in which we expect similar units that only vary on A. This view is summarized by the phrase *fairness given the difference*, which we refer to as the *Kohler-Hausmann Critique* (KHC). We recommend Rose [41] for a comprehensive literature review on the constructivist view.

Remark 3.1 *Under KHC, the ST approach [45, 49] is wrong as it centers both control and test groups on $(\mathbf{x}_c, a_c, \widehat{y}_c)$, failing to account for the systematic effects of A and leading to an unfair counterfactual representation of the complainant c in the form of the constructed test group.*

We address the previous remark in two ways. First, we perform separate searches for the control and test groups of each c by using separate search-centers (next section). Second, we expect a case of indirect discrimination where A does not affect $\widehat{Y}$ directly, but still hinders $\widehat{Y}$ through $\mathbf{X}$. We

[2] Karimi et al. [25] add "structural" to distinguish between counterfactuals generated using SCM from counterfactual explanations [46]. In this paper, counterfactuals are always structural.

formalize this into a SCM $\mathcal{M}$ with DAG $\mathcal{G}$ that describes a *data generating model* (DGM) for $\mathcal{D}$, in which A can only be a parent of $\mathbf{X}$. See Figure 1 as an example. Together, *we insure that the control and test groups are built on different embodiments of "similarity" to c.* For the control group, this is straightforward and in line with standard ST practices: we search and match for similar individuals based on what we observe about c. For the test group, though, we first intervene A and let its effects trickle-down onto $\mathbf{X}$ before we search and match for similar individuals based on a hypothetical, counterfactual version of c. We thus end up comparing control and test groups of not similar individuals.

3.2 Two search-centers for the same complainant

We center the search for the control group around the complainant's *factual tuple*, $(\mathbf{x}_c, a_c, \widehat{y}_c)$, which is just the tuple of interest, while we center the search for the test group around the generated *counterfactual tuple* $(\mathbf{x}_c^{CF}, a_c^{CF}, \widehat{y}_c^{CF})$. Note that the latter is not in $\mathcal{D}$: it belongs to the *counterfactual dataset* $\mathcal{D}^{CF}$ that we generate using the steps introduced in Section 2. We emphasize that the counterfactual tuple does not need to have a different decision outcome than its factual tuple.

Definition 3.1 (The counterfactual dataset) *For a given decision-maker b and known causal graph $\mathcal{G}$, the counterfactual dataset $\mathcal{D}^{CF}$ is the counterfactual mapping of the dataset $\mathcal{D}$ via the abduction, action, and prediction steps [35] based on intervening the protected attribute A from its protected to non-protected values, or $do(A := 0)$.*

All counterfactual)tuples in $\mathcal{D}^{CF}$ are non-protected. This means that all non-protected individuals in $\mathcal{D}_0 \subseteq \mathcal{D}$ keep the same $\mathbf{X}^{CF}$ values. Conversely, this is not true for the protected individuals in $\mathcal{D}_1 \subseteq \mathcal{D}$. Hence, for each complainant c we draw the factual tuple from $\mathcal{D}_1$ (*what is*) and the corresponding counterfactual tuple from $\mathcal{D}_1^{CF}$ (*what would have been if*), such that $\mathcal{D}_1^{CF} \equiv \{l \in \mathcal{D}^{CF} : l \in \mathcal{D}_1\}$, which we call the *counterfactual dataset of complainants*. This means, as the individual identifiers (or row indices) do not change when mapping the counterfactual dataset, that we can trace the factual and counterfactual tuples using c on both datasets.

Causal sufficiency is central to generating $\mathcal{D}^{CF}$. We can relax this assumption by assuming a local, unobserved confounder that can be inferred and controlled for using $\mathcal{D}$ and its associated SCM $\mathcal{M}$. For instance, Kusner et al. [29] use a Monte Carlo Markov Chain while Sánchez-Martín et al. [43] use a variational graph autoencoder to draw the hidden confounder from the available data and restore causal sufficiency. Overall, counterfactual generation is inescapably linked to good information about $\mathcal{G}$ and assumptions about it are difficult to avoid [24].

3.3 Distance function

We use the distance function between two tuples, $d(t, t')$, as defined in k-NN ST [45]. This distance function averages the sum of the per-relevant-attribute distances across $\mathbf{X}$, in which for a continuous, ordinal, or interval x_i attribute we use the normalized Manhattan distance; and for a categorical x_i attribute we use the overlap measurement. The normalized *Manhattan distance* is defines as:

$$md(x_i, x_{i'}) = \frac{|x_i - x_{i'}|}{(\max(X) - \min(X))} \tag{3}$$

and the *overlap measurement* is defined as:

$$ol(x_i, x_{i'}) = \begin{cases} 1, & \text{if } x_i = x_{i'} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

that combined result in the *distance between two tuples* as:

$$d(t, t') = \frac{\sum_{i=1}^{|\mathbf{X}|} d_i(x_i - x_{i'})}{|\mathbf{X}|} \tag{5}$$

where d_i equals the overlap measurement (ol) if the attribute is categorical; otherwise, it equals the Manhattan distance (md). A lower (5) implies a higher similarity between the tuples t and t'. The choice of distance function is not restrictive, and indeed we can explore and compare other options (e.g., heterogeneous distance functions [48]) later on.

3.4 Control and test k-neighborhoods

We use k-NN to construct the control and test groups for k *neighbors*.[3] For the *control group*, we use the factual tuple $(\mathbf{x}_c, a_c, \widehat{y}_c) \in \mathcal{D}_1 \subseteq \mathcal{D}$ as center and search over $\mathcal{D}_1 \subseteq \mathcal{D}$:

$$\textit{k-ctr} \equiv kset_{\mathcal{D}_1}(\mathbf{x}_c, \mathbf{x}_i) = \{\mathbf{x}_i \in \mathcal{D}_1 : rank_{\mathcal{D}_1}(\mathbf{x}_c, \mathbf{x}_i) \leq k\} \tag{6}$$

where $rank_{\mathcal{D}_1}(\mathbf{x}_c, \mathbf{x}_i)$ is the set of i tuples in $\mathcal{D}_1$ ordered increasingly based on $d(\mathbf{x}_c, \mathbf{x}_i)$. Similarly, we use for the *test group* the generated counterfactual tuple $(\mathbf{x}_c^{CF}, a_c^{CF}, \widehat{y}_c^{CF}) \in \mathcal{D}_1^{CF} \subseteq \mathcal{D}^{CF}$ as center and search over $\mathcal{D}_0 \subseteq \mathcal{D}$:

$$\textit{k-tst} \equiv kset_{\mathcal{D}_0}(\mathbf{x}_c^{CF}, \mathbf{x}_i) = \{\mathbf{x}_i \in \mathcal{D}_0 : rank_{\mathcal{D}_0}(\mathbf{x}_c^{CF}, \mathbf{x}_i) \leq k\} \tag{7}$$

Both (6) and (7) are defined per each c and can be expanded, for instance, by including additional constraints, such as a maximum allowed distance $\epsilon > 0$.[4] Notice that here A is not used, only $\mathbf{X}$.

3.5 Detecting discrimination

We compare the control and test groups of each c to detect for indirect individual discrimination. We summarize the information in each group by calculating the *proportion of rejected tuples*:

$$p_1 = \frac{\sum_{i \,\in\, \textit{k-ctr}}^{|\textit{k-ctr}|}(\mathbf{x}_i, a_i = 1, \widehat{y}_i = 0)}{|\textit{k-ctr}|} \qquad\qquad p_2 = \frac{\sum_{i \,\in\, \textit{k-tst}}^{|\textit{k-tst}|}(\mathbf{x}_i, a_i = 0, \widehat{y}_i = 0)}{|\textit{k-tst}|} \tag{8}$$

where *k-ctr* includes the factual and *k-ctr* the counterfactual tuples. Our focus on $\widehat{Y} = 0$ is based on the fact that the individual complainant, in principle, will only "complain" if faced with a negative decision outcome. However, we also consider potential complainants in $\mathcal{D}$ with a positive decision outcome as we are interested in exploring whether this too is the case for its control and test neighbors: e.g., a form of tokenism discrimination [39].

To test for individual discrimination or, at least, for evidence of unfair treatment against a complainant c we compare p_1 and p_2 in (8) using *the difference in the proportions*: $p_1 - p_2$. A positive difference implies that the protected group gets rejected at a higher rate than the non-protected group (potential discrimination); a negative difference, conversely, implies that the non-protected group gets rejected at a higher rate than the protected group (potential reverse discrimination); and a zero difference implies a potentially neutral decision process.

We use "potential" here as there is a chance that the observed $p_1 - p_2$ value for c is due to randomness within the decision process. We follow Thanh et al. [45] and construct the *Wald confidence intervals* (CIs) $[(p_1 - p_2) - w_\alpha, (p_1 - p_2) + w_\alpha]$ for a significance level of $100(\alpha)\%$:

$$w_\alpha = z_{(1-\alpha/2)} \sqrt{\frac{p_1(1-p_1)}{|\textit{k-ctr}|} + \frac{p_2(1-p_2)}{|\textit{k-tst}|}} \tag{9}$$

where z is the $(1 - \alpha/2)$ quantile of the standard normal. We look for a positive deviation from 0, $p_1 - p_2 > 0$, as proof for individual discrimination toward c. This condition implies that the CIs should exclude zero. We can extend this condition by including $\tau \in \mathbb{R}^+$ to denote some desired deviation from zero that we wish to consider to claim discrimination.

Definition 3.2 (Testing individual discrimination claim) *We define discrimination toward complainant c as a strictly positive difference between the rejection rate in the control group, p_1, and the rejection rate in the test group, p_2: $p_1 - p_2 > 0$. Further, we say a complainant's discrimination claim is τ-valid if $p_1 - p_2 > \tau$. Furthermore, the claim is said to be statistically τ-valid given a significance level α if the Wald CIs do not contain τ.*

Remark 3.2 (On counterfactual fairness [29]) *The proportions in* (8) *allude to the factual (p_1) and counterfactual (p_2) worlds of the complainant c as described by the SCM $\mathcal{M}$ and causal graph $\mathcal{G}$ of the dataset $\mathcal{D}$. The difference $p_1 - p_2$ is, in turn, an estimator for the counterfactual fairness of the decision model b for individual c, though, now equipped with a measure of uncertainty based on the constructed control and test groups via cfST.*

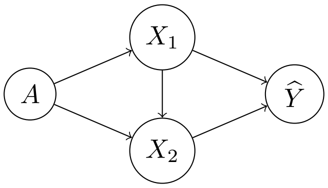
[3] `https://scikit-learn.org/stable/modules/neighbors.html`

[4] For example, *k-ctr* given ϵ: $kset_{\mathcal{D}_1}(\mathbf{x}_c, \mathbf{x}_i) = \{\mathbf{x}_i \in \mathcal{D}_1 : rank_{\mathcal{D}_1}(\mathbf{x}_c, \mathbf{x}_i) \leq k \wedge d(\mathbf{x}_c, \mathbf{x}_i) \leq \epsilon\}$

4 Experiment

We showcase our proposed tool using a synthetic dataset representing the fictitious loan application process illustrated in Figure 1. We test our counterfactual situation testing (cfST) tool and compare it against its standard algorithmic counterpart, the k-nearest neighbors situation testing (k-NN ST) tool [45]. As shown in Table 1, cfST detects more individual discrimination cases than k-NN ST for all k-neighborhood sizes. We also compare our tool against counterfactual fairness (CF) [29] to highlight the latter method's limitations as a tool for detecting discrimination relative to cfST.

Data. We use a modified version of Figure 1 from [25] for our DGM. In our version, Figure 1, we introduce gender as the protected attribute A into the SCM $\mathcal{M}$. Gender directly affects both an individual's annual salary X_1 and bank balance X_2 and indirectly affects the predicted decision outcome $\widehat{Y}$, where $\widehat{Y} = 0$ when the individual loan application is rejected and $\widehat{Y} = 1$ otherwise. We generate $\mathcal{D}$ for $n = 4993$ individuals under $A \sim \text{Uniform}(0.45)$ such that $A = 1$ if the individual is female and $A = 0$ otherwise and assume: $X_1 := (-\$1500) \cdot \text{Poisson}(10) \cdot A + U_1$; $X_2 := (-\$300) \cdot \mathcal{X}^2(4) \cdot A + (3/10) \cdot X_1 + U_2$; and $\widehat{Y} = \mathbf{1}\{X_1 + 5 \cdot X_2 - 225000\}$ with $U_1 \sim \$10000 \cdot \text{Poisson}(10)$ and $U_2 \sim \$2500 \cdot \mathcal{N}(0, 1)$.[5] With A we introduce a systematic bias onto the relevant decision attributes for female applicants. Such "penalties", e.g., could represent the financial burdens female professionals face today in this fictitious world after having been discouraged back when they were students from pursuing high-paying, male-oriented fields (see, e.g., [12]). Under Figure 1, 39.2% of men in $\mathcal{D}$ are rejected for the loan while for women it is 60.9%.



$$\mathcal{M} \begin{cases} A & := U_A \\ X_1 & := f_1(A) + U_1 \\ X_2 & := f_2(X_1, A) + U_2 \end{cases}$$

$$\widehat{Y} = b(X_1, X_2)$$

Figure 1: The data generation model (DGM) with corresponding SCM $\mathcal{M}$ and DAG $\mathcal{G}$ for our experimental synthetic loan application dataset. Let A denote an individual's gender, X_1 annual salary, X_2 bank balance, and $\widehat{Y}$ the loan decision based on the b decision-maker.

For the DGM in Figure 1, we distinguish between the decision-maker b and the SCM $\mathcal{M}$. Implementation-wise, this distinction changes nothing. Conceptually, however, it highlights two aspects relevant to cfST and, overall, discrimination analysis. First, $\mathcal{M}$ is inherit to the individuals while $\widehat{Y}$ is externally imposed by b. It is b that decides what inputs to use for $\widehat{Y}$, not the individuals. Similarly, the individuals cannot modify, assuming no strategic-behaviour [31], their relevant attributes. Second, under cfST, it is the decision-maker b that is put into question by the individual discrimination claim(s). We are not strictly interested here in $\mathbf{X}$, but in what happens to b given $\mathbf{X}$, which results in $\widehat{Y}$ for a given individual.

Implementation. Using cfST, we can test whether a female applicant has been discriminated under b based on what is recorded in $\mathcal{D}$. By implementing cfST, we can detect individual discrimination for each eligible complainant, i.e., female applicant, in $\mathcal{D}$. We can also, in turn, build an overall view on how b behaves toward the group female in this context by aggregating the number of individual discrimination cases detected via cfST. We view this not as a matter of individual versus group fairness metrics, but as a consequence of proving discrimination claims at a larger scale. When proving a discrimination claim, indeed the main focus is on the complainant c and whether there is evidence that c was discriminated against by the decision-maker b. However, in doing so, we also pass judgement onto b: if we have reasons to believe b discriminated against c, then we might have grounds to investigate further the wider claim about b discriminating against the group in which c belongs. The algorithmic setting simply introduces a larger scale: we have more than one potential discrimination claim within $\mathcal{D}$ to test, allowing us to draw individual-level conclusions while motivating future

[5] It was initially $n = 5000$, but after adding A we had to drop those rows with negative X_1 or X_2 values.

group-level ones. This last point motivates our view of cfST as a context-specific algorithmic auditing tool for b that centers on the individual claims gathered in $\mathcal{D}$.

For the dataset $\mathcal{D}$ with DAG $\mathcal{G}$ from Figure 1, we use $\alpha = 5\%$ and $\tau = 0.0$, and assume a set of k requested by a domain (legal) expert. The overall discrimination claim is that the loan decision b, which is known, is unfair toward female applicants. We have a total of 1712 females in the dataset $\mathcal{D}$. Given our starting point for modelling the structural functions, linear Gaussian ANM (Section 2), we estimate $\mathcal{M}$ given the data using the linear regression as it is the best linear unbiased estimate (BLUE) under this setting [17]. Once $\mathcal{M}$ is estimated (the "step 0"), we can generate $\mathcal{D}^{CF}$ by intervening A via $do(A := 0)$, answering the question *what would have been if all applicants were male*. For instance, for the factual tuple $(y = 0, x_1 = 35000, x_2 = 7948, a = 1)$, we find its counterfactual tuple to be $(y^{CF} = 0, x_1^{CF} = 50796, x_2^{CF} = 13852, a^{CF} = 0)$. Figure 2 shows the factual and counterfactual distributions of the relevant attributes for female applicants. As expected, both counterfactual distributions shift to the right (an overall increase in both X_1 and X_2), highlighting the "removal" of the systemic bias imposed on these attributes for being female. The overall loan rejection rate for women drops from 60.9% in $\mathcal{D}$ to 38.7% in $\mathcal{D}^{CF}$, which is now closer to the overall loan rejection rate of 39.2% experienced by men.

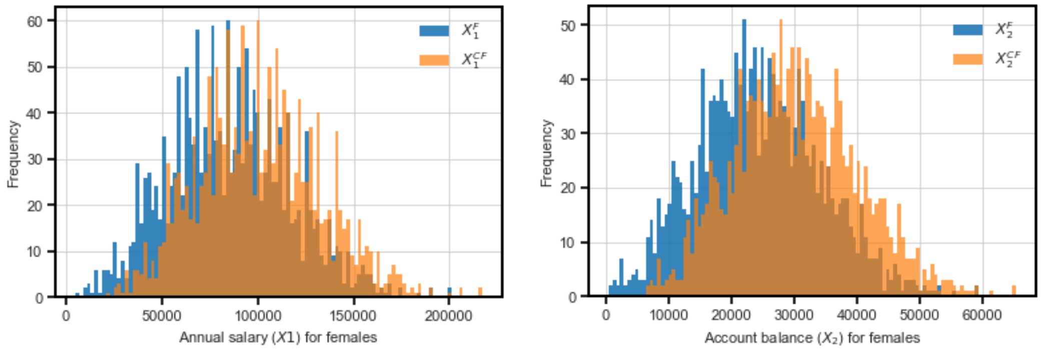


Figure 2: Distributions for annual salary (X_1) and account balance (X_2) for females ($A = 1$) for both factual $\mathcal{D}$ and counterfactual $\mathcal{D}^{CF}$ datasets. For both attributes, there is a rightward shift (an increases) after the intervention $do(A := 0)$.

We run cfST (Algorithm 1) on $\mathcal{D}$ and $\mathcal{D}^{CF}$. Recall from Section 3.2 that we construct the control and test groups for each c complainant using their factual and counterfactual tuples as search centers, respectively, during the k-NN implementation. Using the distance function (5) from Section 3.3, for each c we construct its control group by searching through the protected set $\mathcal{D}_1$ (i.e., the subset of $\mathcal{D}$ containing only female applicants) using the factual tuple in $\mathcal{D}$ while we construct its test group by searching through the non-protected set $\mathcal{D}_0$ (i.e., the subset of $\mathcal{D}$ containing only male applicants) using the generated counterfactual factual tuple in $\mathcal{D}^{CF}$. We present our results in Table 1 for different k-neighborhood sizes, and compare it to k-NN ST. As shown in Table 1, under an $\alpha = 5\%$ and a $\tau = 0.0$, the cfST tool finds a greater number of individual discrimination cases in $\mathcal{D}$ than its k-NN counterpart for all of the indicated k sizes. For instance, for control and test groups of $k = 50$ neighbors, we find 490 cases (or 29% of females in the data) using cfST versus just 84 cases (or 5%) using ST.

Table 1: Number (and %) of individual discrimination cases in $\mathcal{D}$

Method	$k = 0$	$k = 15$	$k = 30$	$k = 50$	$k = 100$
cfST	0	440 (26%)	461 (27%)	490 (29%)	539 (31%)
k-NN ST	0	55 (3.2%)	65 (3.8%)	84 (5%)	107 (6.3%)
CF	376 (22%)	376 (22%)	376 (22%)	376 (22%)	376 (22%)

The results in Table 1 since the k-NN ST is more conservative than the cfST in terms of how it constructs the control and test groups and thus tests for individual discrimination for each c

complainant. The k-NN ST uses a similar algorithm to cfST, with the main exception that the factual tuple is used to search both $\mathcal{D}_1$ and $\mathcal{D}_0$, which ignores how A affects indirectly $\widehat{Y}$ through X_1 and X_2. It performs an *idealized comparison* between what is and what would have been for females in this loan decision process, which is exactly what the KHC criticizes.

Given that we have access to the factual, $\mathcal{D}$, and counterfactual, $\mathcal{D}^{CF}$ worlds, we can judge the counterfactual fairness of b. Based on the CF definition [29], b is counterfactually fair if the factual and counterfactual $\widehat{Y}$'s are the same around the protected attribute A. In our case, b is said to be counterfactually fair if a rejected (or accepted) female applicant would have still been rejected (or accepted) had she been a male applicant. Unsurprisingly, b is counterfactually unfair as it finds 376 cases across $\mathcal{D}$ of rejected female applicants that would have been accepted had they been male applicants. If we used CF as our definition for individual discrimination, these 376 cases of individual discrimination would amount to 22% of females in the data. We present these results also in Table 1 under a scenario without control and test groups (i.e, $k = 0$) as these are not required for the CF method. In turn, this means that the individual discrimination cases detected under a CF are invariant to the size of k as shown in Table 1.

Interpretation. At a high-level, the result in Table 1 show how many individual cases of discrimination each method detects in the dataset $\mathcal{D}$ for the protected group. For instance, at $k = 15$, cfST finds 440 females that were discriminated under Definition 3.2 while k-NN ST 55 and CF 376 under their respective definitions of discrimination. Although these are counts for individual cases, these numbers highlight the bias b has toward female applicants overall.

Regarding cfST versus k-NN ST, the results in Table 1 do not necessarily imply one method is better than the other. Instead, we believe it highlights the implications from changing causal discrimination models when detecting unfair treatment around a protected attribute. The k-NN ST tool performs an idealized comparison for each c to detect individual discrimination, while cfST performs a more flexible comparison by operationalizing "fairness given the difference". More formally, how we construct the control and test groups will depend on the within-group characteristics of the female (or protected) and male (or non-protected) groups that, in turn, will condition the between-group comparison when we compare the rejection rates of each group (recall p_1 and p_2 in (8)). Given the individual focus of ST tools, the choice of the search centers for the control and test groups not only conditions the search (recall (6) and (7)), but it is also a statement on how we conceive the within-group ordering (see, e.g, [10, 9]).

Under the *idealized comparison* performed by k-NN ST, we implicitly assume that the ordering in the female and male groups are the same. For instance, the two tuples $(x_1 = 35000, x_2 = 7948, a = 1)$ and $(x_1 = 35000, x_2 = 7948, a = 0)$ would be the same, meaning these two individuals are similar and could be compared. From an individual fairness [14] perspective, we are saying that it is fair to compare observably similar male and female candidates to test for discrimination. This view, however, as [28] rightfully pointed out and our experimental results illustrate, can miss the pervasive effects of the protected attribute A on the relevant attributes X_1 and X_2. Under the *fairness given the difference* view operationalized by cfST, we do not assume the same within-group ordering across protected and non-protected groups. We would use $(x_1 = 35000, x_2 = 7948, a = 1)$ to compare the individual against other female applicants, while we would use its counterfactual $(x_1^{CF} = 50796, x_2^{CF} = 13852, a^{CF} = 0)$ to compare the individual against other male applicants. Under such counterfactual, the male tuple $(x_1 = 35000, x_2 = 7948, a = 0)$ is no longer similar and thus not a candidate for comparison. This is the basis of cfST. We undertake this approach with the hopes of preserving the within-group orderings, meaning that it is unfair to compare observably similar individuals that do not share the protected group to detect discrimination.

Regarding cfST versus CF, we argue that the results in Table 1 show the limitation of the latter method for detecting discrimination. An aspect that is characteristic of ST [5, 40] and other traditional discrimination analysis tools (e.g., [15, 16, 7]) is the need to construct control and test groups to rule out the possibility of randomness from the decision-making process in question. Under the risk of some randomness, we cannot take the individual comparison literally. The counterfactual query is a sufficient, but not a necessary condition for proving the individual discrimination claim. Under CF, we get an individualized view of the individual discrimination claim by looking only at the factual and counterfactual tuples. Under cfST, we enhance such view by looking at similar individual profiles centered around these two tuples.

5 Discussion

We presented counterfactual situation testing (cfST), a new algorithmic tool for detecting individual discrimination in a dataset. It is a first algorithmic attempt at implementing a new paradigm for addressing fairness problems based on the Kohler-Hausmann Critique (KHC). Under the KHC, we consider the observed differences rather than the similarities between instances to test for the unfair treatment of protected-by-law individuals. For algorithmic fairness in particular, embracing "fairness given the difference" implies formalizing discrimination in a way that is more flexible as similar individuals are no longer expected to be treated similarly within some contexts. It sounds counterintuitive, but it might be a necessary step, modeling-wise, for recognizing the historical processes embodied by A (e.g., race and socioeconomic background in modern USA [44, 42, 2]) and, in turn, for better operationalizing A to address today its systematic effects of the past.

Limitations and next steps. *Real world datasets.* We plan to apply cfST on the revised retiring adult income [13], the revised German credit [18], and the admissions to law school [47] datasets, which are common in the causal fairness literature. For the latter dataset, we also plan to address the issue of handling a hidden confounder by drawing it from the observed dataset $\mathcal{D}$ under a known causal graph [29, 43]. *Other discrimination tools.* Similarly, we plan to compare cfST against the causal ST [49] and the FlipTest [8] methods. *Statistical power analysis.* We would like to derive k using statistical power analysis [11] to decreases the likelihood of a Type-II error when detecting discrimination under cfST. This step would also make the cfST pipeline more data-driven. *Unknown decision-maker b and causal graph $\mathcal{G}$.* Finally, we would like to relax the cfST assumptions of having a known b and $\mathcal{G}$. We can relax these assumption, for instance, by applying a local explainer around the decision boundary in $\mathcal{D}$ [19] and running an causal discovery algorithm on $\mathcal{D}$ [37], respectively. All of these are steps we plan to explore in the immediate future.

Acknowledgments and Disclosure of Funding

This work received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Actions (GA number 860630) for the project "NoBIAS: Artificial Intelligence without Bias".This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

References

[1] J. Adams-Prassl, R. Binns, and A. Kelly-Lyth. Directly discriminatory algorithms. *The Modern Law Review*, 2022.

[2] J. S. Adler. *Murder in New Orleans: the creation of Jim Crow policing*. University of Chicago Press, 2019.

[3] J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 2008.

[4] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104(3): 671–732, 2016.

[5] M. Bendick. Situation testing for employment discrimination in the United States of America. *Horizons stratégiques*, 3(5):17–39, 2007.

[6] M. Bertrand and E. Duflo. Field experiments on discrimination. *Handbook of Economic Field Experiments*, 1:309–393, 2017.

[7] M. Bertrand and S. Mullainathan. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review*, 94(4):991–1013, 2004.

[8] E. Black, S. Yeom, and M. Fredrikson. Fliptest: fairness testing via optimal transport. In *FAT**, pages 111–121. ACM, 2020.

[9] E. Chzhen and N. Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, 2022.

[10] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with wasserstein barycenters. In *NeurIPS*, 2020.

[11] J. Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[12] C. Criado-Perez. *Invisible Women*. Vintage, 2019.

[13] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. In *NeurIPS*, pages 6478–6490, 2021.

[14] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *ITCS*, pages 214–226. ACM, 2012.

[15] M. Fix and R. J. Struyk. *Clear and Convincing Evidence: Measurement of Discrimination in America*. Urban Institute Press, 1993.

[16] C. Goldin and C. Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, September 2000.

[17] W. H. Greene. *Econometric Analysis*. Prentice Hall, 5ht edition, 2002.

[18] U. Groemping. South German credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, 4, 2019.

[19] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intell. Syst.*, 34(6):14–23, 2019.

[20] P. Hacker. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, 55(4), 2018.

[21] J. J. Heckman. Detecting discrimination. *Journal of Economic Perspectives*, 12(2):101–116, 1998.

[22] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, pages 689–696. Curran Associates, Inc., 2008.

[23] L. Hu and I. Kohler-Hausmann. What's sex got to do with machine learning? In *FAT**, page 513. ACM, 2020.

[24] A. Karimi, B. J. von Kügelgen, B. Schölkopf, and I. Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *NeurIPS*, 2020.

[25] A. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *FAccT*, pages 353–362. ACM, 2021.

[26] A. Kasirzadeh and A. Smart. The use and misuse of counterfactuals in ethical machine learning. In *FAccT*, pages 228–236. ACM, 2021.

[27] J. M. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein. Discrimination in the age of algorithms. *CoRR*, abs/1902.03731, 2019. URL `http://arxiv.org/abs/1902.03731`.

[28] I. Kohler-Hausmann. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163, 2018.

[29] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *NIPS*, pages 4066–4076, 2017.

[30] K. Makhlouf, S. Zhioua, and C. Palamidessi. Survey on causal-based machine learning fairness notions. *CoRR*, abs/2010.09553, 2020. URL `https://arxiv.org/abs/2010.09553`.

[31] J. Miller, S. Milli, and M. Hardt. Strategic classification is causal modeling in disguise. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 6917–6926. PMLR, 2020.

[32] A. C. Morgan, S. F. Way, M. J. Hoefer, D. B. Larremore, M. Galesic, and A. Clauset. The unequal impact of parenthood in academia. *Science Advances*, 7(9), 2021.

[33] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[34] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[35] J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

[36] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *KDD*, pages 560–568. ACM, 2008.

[37] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[38] B. Qureshi, F. Kamiran, A. Karim, S. Ruggieri, and D. Pedreschi. Causal inference for social discrimination reasoning. *J. Intell. Inf. Syst.*, 54(2):425–437, 2020.

[39] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.*, 29(5):582–638, 2014.

[40] I. Rorive. Proving discrimination cases: The role of situation testing. *Centre for Equal Rights and MPG*, 2009. URL `https://ec.europa.eu/migrant-integration/library-document/proving-discrimination-cases-role-situation-testing_en`.

[41] E. K. Rose. A Constructivist Perspective on Empirical Discrimination Research. *Working Manuscript*, 2022. URL `https://ekrose.github.io/files/constructivism.pdf`.

[42] R. Rothstein. *The Color of Law: A Forgotten History of How our Government Segregated America*. Liveright Publishing, 2017.

[43] P. Sánchez-Martín, M. Rateike, and I. Valera. VACA: designing variational graph autoencoders for causal queries. In *AAAI*, pages 8159–8168. AAAI Press, 2022.

[44] E. C. Schneider. *Smack: Heroin and the American city*. University of Pennsylvania Press, 2008.

[45] B. L. Thanh, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *KDD*, pages 502–510. ACM, 2011.

[46] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[47] L. F. Wightman. *LSAC national longitudinal bar passage study*. Law School Admission Council, 1998.

[48] D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. *J. Artif. Intell. Res.*, 6:1–34, 1997.

[49] L. Zhang, Y. Wu, and X. Wu. Situation testing-based discrimination discovery: A causal inference approach. In *IJCAI*, pages 2718–2724. IJCAI/AAAI Press, 2016.

A Algorithms for cfST

In this section, we present the relevant pseudo-algorithms for the proposed counterfactual situation testing (cfST) method. Algorithm 1 performs cfST while Algorithm 2 returns the indices of the top-k tuples with respect to the search center based on the distance function from Section 3.3.

Notice that the main difference in Algorithm 1 when creating the neighborhoods is that the centers are drawn from the factual dataset (for the control group) and the counterfactual dataset (for the test group). We use the same c (i.e., index) for both as these two data-frames have the same structure by construction.

Algorithm 1: run_cfST()

```
Input  : $\mathcal{D}, \mathcal{D}^{CF}, k$
Output : $[p_1 - p_2]$
$prot\_condition \leftarrow \mathcal{D}[:,\ prot\_attribute] == prot\_value$
$\mathcal{D}_1 \leftarrow \mathcal{D}[prot\_condition]$                     // get protected (control) search space
$\mathcal{D}_0 \leftarrow \mathcal{D}[\neg\ prot\_condition]$                  // get non-protected (test) search space
$prot\_idx \leftarrow \mathcal{D}_1.index.to\_list(\ )$;                    // get idx for all complainants
$diff\_list = [\ ]$
for $c,\ row \in prot\_idx$ do
    $res\_1 \leftarrow get\_top\_k(\mathcal{D}\quad [c,\ :], \mathcal{D}_1, k)$;      // idx of the top-k tuples for control group
    $res\_2 \leftarrow get\_top\_k(\mathcal{D}^{CF}[c,\ :], \mathcal{D}_0, k)$;       // idx of the top-k tuples for test group
    $p_1 \leftarrow sum(\mathcal{D}[res_1,\ target\_attribute] == negative\_outcome)\ /\ len(res\_1)$
    $p_2 \leftarrow sum(\mathcal{D}[res_2,\ target\_attribute] == negative\_outcome)\ /\ len(res\_2)$
    $diff\_list[c] \leftarrow p_1 - p_2$
end
return $diff\_list$
```

Algorithm 2: get_top_k

```
Input  : $t, t\_set, k$
Output : $[indices]$
$(idx, dist) \leftarrow k\_NN(t, t\_set, k)$;                        // run k-NN algorithm
$remove\_t(idx, dist)$;                                // remove the center t from idx
$(idx', dist') \leftarrow sort(idx, dist)$;                           // sort idx by the distance
return $(idx',\ )$
```