Counterfactual Situation Testing: Uncovering Discrimination under Fairness given the Difference

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

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

ABSTRACT

We present counterfactual situation testing (CST), a causal data mining framework for detecting individual discrimination in a dataset of classifier decisions. CST answers the question "what would have been the model outcome had the individual, or complainant, been of a different protected status?" in an actionable and meaningful way. It extends the legally-grounded situation testing of Thanh et al. [62] by operationalizing the notion of *fairness given the difference* of Kohler-Hausmann [38] using counterfactual reasoning. In standard situation testing we find for each complainant similar protected and non-protected instances in the dataset; construct respectively a control and test group; and compare the groups such that a difference in decision outcomes implies a case of potential individual discrimination. In CST we avoid this idealized comparison by establishing the test group on the complainant's counterfactual generated via the steps of abduction, action, and prediction. The counterfactual reflects how the protected attribute, when changed, affects the other seemingly neutral attributes of the complainant. Under CST we, thus, test for discrimination by comparing similar individuals within each group but dissimilar individuals across both groups for each complainant. Evaluating it on two classification scenarios, CST uncovers a greater number of cases than ST, even when the classifier is counterfactually fair.

CCS CONCEPTS

• **Computing methodologies → Classification**; **Machine learning**.

KEYWORDS

discrimination discovery, structural causal models, counterfactual fairness

1 INTRODUCTION

Automated decision making (ADM) is becoming ubiquitous and its societal discontents clearer [4, 19, 31]. There is a shared urgency by regulators [21, 65] and researchers [37, 57] to develop frameworks that can asses these classifiers for potential discrimination based on protected attributes such as gender, race, or religion. Discrimination is often conceived as a causal claim on the effect of the protected attribute over an individual decision outcome [24, 30]. It is, in particular, a conception based on counterfactual reasoning—what would have been the model outcome if the individual, or *complainant*, were of a different protected status?—where we "manipulate" the protected attribute of the individual. Kohler-Hausmann [38] calls such conceptualization the *counterfactual causal model of discrimination* (CMD).

Several frameworks for proving ADM discrimination are based on CMD [38]. Central to these frameworks is defining "similar" instances to the complainant; arranging them based on their protected status into control and test groups; and comparing the decision outcomes of these groups to detect the effect of the protected attribute. Among the available tools [12, 41, 52], however, there is a need for one that is both *actionable* and *meaningful*. We consider a framework to be actionable if it can rule out random circumstances for the individual discrimination claim as often required by courts (e.g., [23, 24, 45]), and meaningful if it can account for known links between the protected attribute and all other attributes when manipulating the former as often demanded by social scientists (e.g., [11, 34, 61]).

In this paper we present *counterfactual situation testing* (CST), a causal data mining framework for detecting instances of individual discrimination in the dataset used by a classifier. It combines (structural)[1] counterfactuals [46, 47] with situation testing [62, 70]. *Counterfactuals* answer to counterfactual reasoning and are generated via structural causal models. Under the right causal knowledge, counterfactuals reflect at the individual level how changing the protected attribute affects other seemingly neutral attributes of a complainant. *Situation testing* is a data mining method, based on the homonymous legal tool [7, 54]. For each complainant, under some search algorithm and distance function for measuring similarity, it finds and compares a control and test group of similar protected and non-protected instances in the dataset, where a difference between the decision outcomes of the groups implies potential discrimination. CST follows the situation testing pipeline with the important

---

[1]Not to be confused with counterfactual explanations [64] from the XAI literature, which are not based on structural causal models. Other works like [33] use "structural" to differentiate counterfactuals from counterfactual explanations. In this paper, counterfactuals *are* structural.

exception that it constructs the test group around the complainant's counterfactual instead of the complainant.

An illustrative example. Consider the scenario in Fig. 1 (used later in Section 4.1) where a bank uses a classifier to accept or reject ($\hat{Y}$) individual loan applications based on annual salary (X_1) and account balance (X_2). Suppose a female applicant ($A = 1$) with $x_1 = 35000$ and $x_2 = 7048$ gets rejected and files for discrimination. The bank is using non-sensitive information to calculate $\hat{Y}$, but according to Fig. 1 there is also a known link between A and $\{X_1, X_2\}$ that questions the neutrality of such information. Under situation testing, we would find a number of female (protected) and male (non-protected) instances with similar characteristics to the complainant. The resulting control and test groups to be compared would both have similar X_1 and X_2 to the complainant. On one hand, comparing multiple instances allows to check whether the complainant's claim is an isolated event or representative of an unfavorable pattern toward female applicants by the model (i.e., actionability). On the other hand, knowing what we know about A and its influence, would it be fair to compare the similar female and male instances? As argued by previous works [32, 38], the answer is no. This *idealized comparison* takes for granted the effect of gender on annual salary and account balance. Under counterfactual situation testing, instead, we would generate the complainant's counterfactual under the causal knowledge provided, creating a "male" applicant with a higher $x_1 = 50796$ and $x_2 = 13852$, and use it rather than the complainant to find similar male instances. The resulting control and test groups would have different X_1 and X_2 between them. This disparate comparison embodies *fairness given the difference*, explicitly acknowledging the lack of neutrality when looking at X_1 and X_2 based on A (i.e., meaningfulness). Here, the control group represents the observed factual world while the test group the hypothetical counterfactual world of the complainant.

In addition, with counterfactual situation testing we propose an actionable extension to *counterfactual fairness* by Kusner et al. [39], which remains the leading causal fairness framework [41]. A classifier is counterfactually fair when the complainant's and its counterfactual's decision outcomes are the same. These are the same two instances used by CST to construct, respectively, the control and test groups, which allows to equip this fairness definition with measures for uncertainty. Hence, CST links counterfactual fairness claims with notions of statistical significance. Further, by looking at the control and test groups rather than the literal comparison of the factual versus counterfactual instances, CST evaluates whether the counterfactual claim is representative of similar instances. Hence, CST detects cases of individual discrimination that are also counterfactually fair, capturing the realistic scenario where a deployed model tends to discriminate when asked to evaluate a borderline instance multiple times.

Based on two case studies using synthetic and real data, we evaluate the CST framework using a *k*-nearest neighbor implementation, k-NN CST, and compare it to its situation testing counterpart, k-NN ST [62], as well as to counterfactual fairness [39]. Here, *k* denotes the number of instances we wish to find for each control and test groups. The experiments show that CST detects a higher number of individual cases of discrimination across the different *k* sizes. Further, the results also show that individual discrimination can

occur even when the model is counterfactually fair. The results hold when dealing with multiple protected attributes as well as different implementation parameters.

Our main contributions with CST are: *(1)* a meaningful and actionable framework for detecting individual discrimination; *(2)* a first operationalization of *fairness given the difference* for discrimination analysis; and *(3)* an actionable extension of counterfactual fairness equipped with confidence intervals. With this in mind, Section 2 explores the role of causal knowledge in CST. Section 3 presents the CST framework and its k-NN implementation, while Section 4 showcases CST via two classification scenarios. Section 5 concludes the paper.

1.1 Related Work

We position CST with current works along the goals of actionability and meaningfulness. Regarding actionability, when proving discrimination, it is important to insure that the framework accounts for sources of randomness in the decision process. Popular non-algorithmic frameworks—such as natural [26] and field [8] experiments, audit [22] and correspondence [9, 53] studies—address this issue by using multiple observations to build inferential statistics. Similar statistics are sometimes asked in court for proving discrimination (e.g., [23, Section 6.3]). Few algorithmic frameworks, instead, address this issue due to model complexity preventing formal inference [5]. An exception are data mining frameworks for discrimination discovery [48, 58] that operationalize the non-algorithmic notions, including situation testing [62, 70]. These frameworks (e.g., [2, 25, 51]) keep the focus on comparing multiple control-test instances for making individual claims, providing evidence similar to that produced by the quantitative tools used in court [38]. To the best of our knowledge, it remains unclear if the same can be said about existing causal fair machine learning methods [41] as these have yet to be used beyond academic circles.

Regarding meaningfulness, situation testing and the other methods have been criticized for their handling of the counterfactual question behind the causal model of discrimination [32, 34, 38]. In particular, these actionable methods take for granted the influence of the protected attribute on all other attributes. This can be seen, e.g., in how situation testing constructs the test group, which is equivalent to changing the protected attribute while keeping everything else equal. Such approach goes against how most social scientists interpret the protected attribute and its role as a social construct when proving discrimination [11, 28, 55, 61]. It is in that regard where structural causal models [46] and their ability for conceiving counterfactuals (e.g., [13, 69]), including counterfactual fairness [39], have an advantage. What the criticisms on counterfactuals [32, 34] overlook here is that generating counterfactuals, as long as the causal knowledge is properly specified, accounts modeling-wise for the effects of changing the protected attribute on all other observed attributes. A framework like counterfactual fairness, relative to situation testing and these other methods, is more meaningful in its handling of protected attributes. The novelty in CST is bridging these two lines of work, borrowing the actionability aspects from situation testing and the meaningful aspects from counterfactual fairness.

2 CAUSAL KNOWLEDGE FOR DISCRIMINATION

Counterfactual situation testing requires access to the dataset of decision records of interest, $\mathcal{D}$, and the algorithmic decision-maker that produced it, $b()$. Let $\mathcal{D}$ contain the set of relevant attributes X, the set of protected attributes A, and the decision outcome $\hat{Y} = b(X)$. We describe $\mathcal{D}$ as a collection of n tuples, each $(x_i, a_i, \hat{y}_i)$ representing the ith individual profile, with $i \in [1, n]$. $\hat{Y}$ is binary with $\hat{Y} = 1$ denoting the positive outcome (e.g., loan granted). For illustrative purposes, we assume a single binary A with $A = 1$ denoting the protected status (e.g., female), though we relax this assumption in the experiments of Section 4.2. We also require causal knowledge in the form of a structural causal model that describes the data generating model behind $\mathcal{D}$. We view this requirement as an input space for experts as these models are a convenient way for organizing assumptions on the source of the discrimination, facilitating stakeholder participation and supporting collaborative reasoning about contested concepts [44].

2.1 Structural Causal Models and Counterfactuals

A *structural causal model* (SCM) [46] $\mathcal{M} = \{\mathcal{S}, \mathcal{P}_{\mathbf{U}}\}$ describes how the set of p variables $W = X \cup A$ is determined based on corresponding sets of structural equations $\mathcal{S}$ and latent variables U with prior distribution $\mathcal{P}_{\mathbf{U}}$. Each $W_j \in W$ is assigned a value through a deterministic function $f_j \in \mathcal{S}$ of its causal parents $W_{pa(j)} \subseteq W \setminus \{W_j\}$ and noise variable U_j with distribution $P(U_j) \in \mathcal{P}_{\mathbf{U}}$. Formally, for $W_j \in W$ we have that $W_j \leftarrow f_j(W_{pa(j)}, U_j)$, indicating the flow of information in terms of child-parent or cause-effect pairs. We consider the associated *causal graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where a node $V_j \in \mathcal{V}$ represents a W_j variable and a directed edge $E_{(j,j')} \in \mathcal{E}$ a causal relation.

We make two assumptions on $\mathcal{M}$ common within the causal fairness literature [41]. First, we assume *causal sufficiency*, meaning there are no hidden common causes in $\mathcal{M}$, or confounders. Second, we assume $\mathcal{G}$ to be *acyclic*, which turns $\mathcal{G}$ into a directed acyclical graph (DAG), allowing for no feedback loops. We write $\mathcal{M}$ under these assumptions as:

$$\mathcal{M} = (\mathcal{S}, \mathcal{P}_U), \quad \mathcal{S} = \{W_j \leftarrow f_j(W_{pa(j)}, U_j)\}_{j=1}^{p},$$
$$\mathcal{P}_{\mathbf{U}} = P(U_1) \times \cdots \times P(U_p) \tag{1}$$

where these assumptions are necessary for generating counterfactuals. The causal sufficiency assumption is particularly deceitful as it is difficult to both test and account for a hidden confounder [18, 40, 43]. The risk of a hidden confounder is a general problem to modeling fairness. Here, the dataset $\mathcal{D}$ delimits our context. We expect it to contain all relevant information used by the decision-maker $b()$.

Input from several stakeholders is needed to derive (1). We see it as a necessary collaborative effort: *before we implement CST, we first need to agree over a worldview for the discrimination context.* Based on $\mathcal{D}$, a domain-expert motivates a causal graph $\mathcal{G}$. A modelling-expert then translates this graphical information into a SCM $\mathcal{M}$, making model specification decisions on $\mathcal{S}$. We do not cover this process here, but this is how we envision the initial implementation stage of counterfactual situation testing.

For a given SCM $\mathcal{M}$ we want to run *counterfactual queries* to build the test group for a complainant. Counterfactual queries answer to *what would have been if* questions. In CST, we wish to ask such questions around the protected attribute A, by setting A to the non-protected status α using the *do-operator* $do(A := \alpha)$ [46] to capture the individual-level effects A has on X according to (1). Let X^{CF} denote the set of counterfactual variables obtained via the three step procedure by Pearl et al. [47]. *Abduction*: for each prior distribution $P(U_i)$ that describes U_i, we compute its posterior distribution given the evidence, or $P(U_i \mid X, A)$. *Action*: we intervene A by changing its structural equation to $A := \alpha$, which gives way to a new SCM $\mathcal{M}'$. *Prediction*: we generate the *counterfactual distribution* $P(X_{A \leftarrow \alpha}^{CF}(U) \mid X, A)$ by propagating the abducted $P(U_i \mid X, A)$ through the revised structural equations in $\mathcal{M}'$.

Note that generating counterfactuals and, thus, CST, unlike, e.g., counterfactual explanations [64] and discrimination frameworks like the FlipTest [10], does not require a change in the individual decision outcome. It is possible for $\hat{Y} = \hat{Y}^{CF}$ after manipulating A. We include a working example on generating counterfactuals in the Appendix.

2.2 Conceiving Discrimination

The legal setting of interest is indirect discrimination under EU law. It occurs when an apparently neutral practice disadvantages individuals that belong to a protected group. Following [27], we focus on indirect discrimination for three reasons. First, unlike disparate impact under US law [6], the decision-maker can still be liable for it despite lack of premeditation and, thus, all practices need to consider potential indirect discrimination implications. Second, many ADM models are not allowed to use the protected attribute as input, making it difficult for regulators to use the direct discrimination setting. Third, we conceive discrimination as a product of a biased society where $b()$ continues to perpetuate the bias reflected in $\mathcal{D}$ because it cannot escape making a decision based on X to derive $\hat{Y}$.

We view the indirect setting as the one that best describes how biased information can still be an issue for an ADM that never uses the protected attribute. Previous causal works [13, 36, 50] have focused more on whether the paths between A and $\hat{Y}$ are direct or indirect. Here, the causal setting is much simpler. We know that $b()$ only uses X, and are more interested in how information from A is carried by X and how can we account for these links using causal knowledge. That said, this does not mean that CST cannot be implemented in other discrimination settings. We simply acknowledge that it was developed with the EU legal framework in mind. Proxy discrimination [63], e.g., is one setting that overlaps with the one we have considered.

Finally, we note that an open legal concern for CST is detecting algorithmic discrimination for various protected attributes, or $|A| > 1$. Two kinds of discrimination, *multiple* and *intersectional*, can occur. Consider, e.g., a black female as the complainant. On what protected attribute is she being potentially discriminated on? In multiple discrimination, we would need to detect separately whether the complainant was discriminated as a black and female individual. In intersectional discrimination, we would instead need to detect simultaneously if the complainant was discriminated as a black-female individual. Only multiple discrimination is currently

recognized by EU law, which is an issue as an individual can be free from multiple discrimination but fall victim of intersectional discrimination [68]. CST can account operationally for both scenarios.

2.3 Fairness given the Difference: the Kohler-Hausmann Critique

Here, we make the case—very briefly—that the causal knowledge required for CST makes it a meaningful framework with respect to situation testing [62, 70] and other tools [52] for detecting discrimination. The reference work is Kohler-Hausmann [38]. We refer to the phrase *fairness given the difference*,[2] which best captures her overall critique toward the causal model of discrimination, as the *Kohler-Hausmann Critique* (KHC). CST aims to be meaningful by operationalizing the KHC. It builds the test group on the complainant's counterfactual, letting X^{CF} reflect the effects of changing A instead of assuming $X = X^{CF}$. This is because we view the test group as a representation of the hypothetical counterfactual world of the complainant.

As argued by [38] and others before [11, 61], it is difficult to deny that most protected attributes, if not all of them, are *social constructs*. That is, these attributes were used to classify *and* divide groups of people in a systematic way that conditioned the material opportunities of multiple generations [42, 55]. Thus, *recognizing A as a social constructs means recognizing that its effects can be reflected in seemingly neutral variables in X*. It is recognizing that A, the attribute, cannot capture alone the meaning of belonging to A and that we might, as a minimum, have to link it with other attributes to better capture this, such as $A \rightarrow X$ where A and X change in unison. These attributes *summarize the historical processes that fairness researchers are trying to address* today and should not be treated lightly.[3]

The notion of *fairness given the difference* centers on how A is treated in the counterfactual causal model of discrimination (CM). The critique goes beyond the standard manipulation concern [3] in which A is an inmutable attribute. Instead, granted that we *can* or, more precisely, *have to* manipulate A for running a discrimination analysis, the critique goes against how most discrimination frameworks operationalize such manipulation. The KCH emphasizes that *when A changes, X should change as well*.

Based on KHC, we consider two types of manipulations that summarize existing frameworks. The *ceteris paribus* (CP), or all else equal, manipulation in which A changes but X remains the same. Examples of it include situation testing [62, 70] but also, e.g., the famous correspondence study by Bertrand and Mullainathan [9]. The *mutatis mutandis* (MM), or changing what needs to be changed, manipulation in which X changes when we manipulate A based on some additional knowledge, like a structural causal model, that explicitly links A to X. Counterfactual fairness [39], e.g., uses this manipulation. The MM is clearly preferred over the CP manipulation when we view A as a social construct.

---

[2]A phrase by Kohler-Hausmann during a panel discussion at NeurIPS 2021 workshop on 'Algorithmic Fairness through the Lens of Causal Reasoning.
[3]A clear example of this would be the use of race by US policy makers during the early post-WWII era. See, e.g., the historical evidence provided by Rothstein [56] (for housing), Schneider [60] (for narcotics), and Adler [1] (for policing).

3 COUNTERFACTUAL SITUATION TESTING

The objective of CST is to *construct* and *compare* a control and test group for each c protected individual, or *complainant*, in $\mathcal{D}$ in a meaningful and actionable way. Let $(x_c, a_c, \widehat{y_c}) \in \mathcal{D}$ denote the *tuple of interest* on which the individual discrimination claim focuses on, where $c \in [1, n]$. We assume access to the ADM $b()$, the dataset $\mathcal{D}$, and a structural causal model $\mathcal{M}$ describing the discrimination context.

There are three key inputs to consider: the *number of instances per group*, k; the *similarity distance function of choice*, d; and the *strength of the evidence for rejecting the discrimination claim*, α. A fourth key input that we fix in this paper is the *search algorithm of choice*, ϕ, which we set as the *k-nearest neighbors algorithm* (k-NN) [29]. We do so as the k-NN is intuitive, easy to implement, and commonly used by existing situation testing frameworks. We discuss the CST implementation used in Section 3.4, including the choice of d.

3.1 Building Control and Test Groups

For complainant c, the control and test groups are built on the *search spaces* and *search centers* for each group. The search spaces are derived and, thus, delimited by $\mathcal{D}$: we are looking for individuals that have gone through the same decision process as the complainant. The search centers, however, are derived separately: the one for the control group comes from $\mathcal{D}$, while the one for the test group comes from the corresponding *counterfactual dataset* $\mathcal{D}^{CF}$. The test search center represents the *what would have been if* of the complainant under a *mutatis mutandis* (MM) manipulation of the protected attribute A that motivates the discrimination claim.

Definition 3.1 (Search Spaces). Under a binary A, where $A = 1$ denotes the protected status, we partition $\mathcal{D}$ into the *control search space* $\mathcal{D}_c = \{(x_i, a_i, \widehat{y_i}) \in \mathcal{D} : a_i = 1\}$ and the *test search space* $\mathcal{D}_t = \{(x_i, a_i, \widehat{y_i}) \in \mathcal{D} : a_i = 0\}$.

Definition 3.2 (Counterfactual Dataset). The counterfactual dataset $\mathcal{D}^{CF}$ represents the counterfactual mapping of each instance in the dataset $\mathcal{D}$, with known decision maker $b()$ and SCM $\mathcal{M}$, via the abduction, action, and prediction steps [47] when setting a binary A to the non-protected value, or $do(A := 0)$.

To obtain $\mathcal{D}^{CF}$, we consider an SCM $\mathcal{M}$, as in (1), where A has no causal parents, or is a root node, A affects only the elements of X considered by the expert(s), and $\hat{Y} = b(X)$. Therefore, when generating the counterfactuals on A (Section 2.1), under the indirect discrimination setting (Section 2.2), the resulting X^{CF} in $\mathcal{D}^{CF}$ should reflect an MM manipulation (Section 2.3). Under this structural representation, if A changes then X changes too. See, e.g., Fig. 1 and Fig. 3. The counterfactual dataset represents the world that the complainants would have experienced under $A = 0$ *given our worldview*. All three definitions extend to $|A| > 1$.

Definition 3.3 (Search Centers). For a complainant c, we use x_c from the tuple of interest $(x_c, a_c, \hat{y}_c) \in \mathcal{D}$ as the control search center for exploring $\mathcal{D}_c \subset \mathcal{D}$, and use x_c^{CF} from the tuple of interest's generated counterfactual $(x_c^{CF}, a_c^{CF}, \hat{y}_c^{CF}) \in \mathcal{D}^{CF}$ as the test search center for exploring $\mathcal{D}_t \subset \mathcal{D}$.

Given the factual $\mathcal{D}$ and counterfactual $\mathcal{D}^{CF}$ datasets, we construct the control and test groups for c using the k-NN algorithm under some distance function $d(x, x')$ to measure similarity between two tuples x and x'. We want each group or neighborhood to have a size k. For the *control group* (k-ctr) we use the (factual) tuple of interest $(x_c, a_c, \hat{y}_c) \in \mathcal{D}$ as search center to explore the protected search space $\mathcal{D}_c$:

$$k\text{-}ctr = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D}_c : rank_d(x_c, x_i) \leq k\} \tag{2}$$

where $rank_d(x_c, x_i)$ is the rank position of x_i among tuples in $\mathcal{D}_c$ with respect to the ascending distance d from x_c^{CF}. For the *test group* (k-tst) we use the counterfactual tuple of interest $(x_c^{CF}, a_c^{CF}, \widehat{y}_c^{CF}) \in \mathcal{D}^{CF}$ as search center to explore the non-protected search space $\mathcal{D}_t$:

$$k\text{-}tst = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D}_t : rank_d(x_c^{CF}, x_i) \leq k\} \tag{3}$$

where $rank_d(x_c^{CF}, x_i)$ is the rank position of x_i among tuples in $\mathcal{D}_t$ with respect to the ascending distance d from x_c^{CF}. We use the same distance function d for each group. Neither A nor $\hat{Y}$ are used for constructing the groups. Both (2) and (3) can be expanded by including additional constraints, such as a maximum allowed distance.

The choice of search centers (Def. 3.3) is what operationalizes *fairness given the difference* for counterfactual situation testing, making it a *meaningful framework* for testing individual discrimination. To build *k-ctr* and *k-tst* using, respectively, x_c and x_c^{CF} is a statement on how we perceive *within group ordering* as imposed by the protected attribute A. This is because the search centers must reflect the A-specific ordering of the search spaces that each center targets. Let us consider our illustrative example from Section 1. If being a female ($A = 1$) in this society imposes certain systematic limitations that hinder x_c, then comparing c to other female instances in the protected search space preserves the group ordering prescribed by $X|A = 1$ as all instances involved experience A in the same way. Therefore, given our worldview, the generated counterfactual male instance for c should then reflect the group ordering prescribed by $X|A = 0$. We expect $x_c \neq x_c^{CF}$ given what we know about the effects of A on X. Using x_c^{CF} as the test search center would allow us to compare c to other male tuples in the non-protected search space without having to reduce A to a phenotype.

One way to look at the previous statement is by considering the notion of effort. If being female requires a higher individual effort to achieve the same x_c, then it is fair to compare c to other female instances. However, it is unfair to compare c to other male instances without adjusting for the extra effort not incurred by the male instances for being males. The counterfactual x_c^{CF} should reflect said adjustment. See [14, 15] on a similar, more formal critique on individual fairness [20] notions.

3.2 Detecting Discrimination

For a complainant c, we compare the control and test groups by looking at the *difference in proportion of negative decision outcomes*, or $\Delta p = p_c - p_t$, such that:

$$p_c = \frac{|\{(x_i, a_i, \widehat{y}_i) \in k\text{-}ctr : \hat{y}_i = 0\}|}{k}$$
$$p_t = \frac{|\{(x_i, a_i, \widehat{y}_i) \in k\text{-}tst : \hat{y}_i = 0\}|}{k} \tag{4}$$

where p_c and p_t represents the count of tuples with a negative decision outcome ($\hat{Y} = 0$), respectively, in the control and test group. Note that only $\hat{Y}$ is used for deriving the proportions. We compute Δp for all protected tuples in $\mathcal{D}$ regardless of the their decision outcome $\hat{Y}$.

CST has the option to include or exclude the search centers when calculating (4). If we exclude them, then p_c and p_t remain as is; if we include them, then $\hat{y}_c$ and $\hat{y}_c^{CF}$ are counted in p_c and p_t, leading to a denominator in both of $k + 1$. We add this option to be able to compare CST against standard situation testing [62, 70], which excludes the search centers, and counterfactual fairness [39], which only uses the search centers.

Since Δp is a proportion comparison, it is known to be *asymptotically normally distributed*, which allows to build *Wald confidence intervals* (CI) around it [62]. Let $z_{\alpha/2}$ be the $1 - \alpha/2$ quantile of the standard normal distribution $\mathcal{N}$ for a *significance level* of α (or, conversely, a *confidence level* $(1 - \alpha) \cdot 100\%$). We write the two-sided CI for Δp of c as:

$$[\Delta p - w_\alpha, \Delta p + w_\alpha], \quad \text{with} \quad w_\alpha = z_{\alpha/2} \sqrt{\frac{p_c(1 - p_c) - p_t(1 - p_t)}{k}} \tag{5}$$

The confidence interval (5) responds to the hypothesis that there is individual discrimination, providing a measure of certainty on Δp through a range of possible values. For a given claim, if the CI contains the *minimum accepted deviation* τ, we cannot reject the hypothesis of no discrimination with $(1 - \alpha) \cdot 100\%$ confidence. In other words, the *null hypothesis* $H_0 : \pi = \tau$ cannot be rejected in favor of the *alternative hypothesis* $H_1 : \pi > \tau$, where π is the true difference in proportion of negative decision outcomes. τ, with a default choice of $\tau = 0$, represents the minimum amount of difference between p_c and p_t that we need to observe to claim individual discrimination. The overall choice of α and τ will depend on the context of the discrimination claim. It can be motivated, for instance, by legal requirements (set, e.g., by the court [62]), or technical requirements (set, e.g., via power analysis [16]), or both.

Definition 3.4 (Individual Discrimination). There is potential[4] individual discrimination toward the complainant c if $\Delta p = p_c - p_t > \tau$, meaning the negative decision outcomes rate for the control group is greater than for the test group by some minimum deviation $\tau \in \mathbb{R}^+$.

We do not view Def. 3.4 as a matter of individual versus group fairness. When we test whether $b()$ discriminates against c, we inevitably pass judgement onto the classifier $b()$ in fear that this behaviour has happened before. In $\mathcal{D}$ we have more than one potential discrimination claim to consider under CST, allowing to draw individual-level conclusions while motivating group-level ones. If $b()$ discriminated against c, it also discriminated against what c represents in terms of membership to A.

Definition 3.5 (Confidence on the Individual Discrimination Claim). The Wald CI (5) gives a measure of certainty on Δp, which is (asymptotically) normally distributed. For a significance level α, we are $(1 - \alpha)\%$ confident on Δp. The claim is said to be statistically valid

---

[4]Or *prima facie*. In practice, discrimination needs to be argued against/for. CST alone, as with any other discrimination analysis tool, cannot claim to prove discrimination. It can, however, provide evidence against/for a discrimination case [52].

if the Wald CI excludes τ. This is a statistical inference extension of Def. 3.4.

The many-to-many comparison behind Δp is what makes counterfactual situation testing an *actionable framework* for testing individual discrimination. Here, the notion of repetition and its relation to representativeness and certainty concerns is important. For proving individual discrimination a single comparison is not enough [23, Sec. 6.3]. This is because we want to ensure, one, that the individual claim is representative of the population, and two, be certain about the individual claim. Implicit to both concerns is finding a pattern of unfavorable decisions against the protected group to which the individual complainant belongs to, i.e., discrimination.

Ideally, we would repeat the decision process multiple times for the discriminatory pattern to become apparent. This is not possible in practice. Back to our illustrative example from Section 1, we cannot ask the female complainant to apply multiple times to the same bank. We instead can look at other similar instances under the same process. This is what p_c and p_t (4) and Def. 3.4 represent. Similarly, if the bank's $b()$ is shown to discriminate against the female complainant, what rules out that it has not done it before or that this one time was an exception? Again, we cannot repeat the decision process until we are certain of the individual discrimination claim. We instead can assume a theoretical distribution of comparisons with π to account for potential randomness in what we detect from the single point estimate that is Δp. This is what the CI (5) and Def. 3.5 represent.

3.3 Connection to Counterfactual Fairness

There is a clear link between CST and *counterfactual fairness* [39]. A decision maker is counterfactually fair if it outputs the same outcome for the factual tuple as for its counterfactual tuple, where the latter is generated based on the abduction, action, and prediction steps and the intervention on the protected attribute.[5] The factual $(x_c, a_c, \hat{y}_c)$ and counterfactual $(x_c^{CF}, a_c^{CF}, \hat{y}_c^{CF})$ tuples for c used in CST are also the ones used for counterfactual fairness. We view CST, when including the search centers, as an actionable extension of counterfactual fairness.

PROPOSITION 3.6 (ON ACTIONABLE COUNTERFACTUAL FAIRNESS). *Counterfactual fairness does not imply nor it is implied by Individual Discrimination (Def. 3.4).*

We present a sketch of proof to Prop. 3.6 the in Appendix. Intuitively, it is possible to handle *borderline cases* where the tuple of interest and its counterfactual both get rejected by $b()$, though the latter is close to the decision boundary. The model $b()$ would be considered counterfactually fair, but would that disprove the individual discrimination claim? CST, by constructing the control and test groups around this single comparison, accounts for this actionability concern. CST further equips counterfactual fairness with confidence intervals. Previous works have addressed uncertainty in counterfactual fairness [35, 59], but with a focus on the structure of the SCM $\mathcal{M}$. We instead address certainty on the literal comparison that motivates the counterfactual fairness definition.

3.4 An Implementation: k-NN CST

Finally, we propose an implementation to our counterfactual situation testing framework. We already defined the search algorithm ϕ as the k-NN algorithm. We define as the similarity measure d the same distance function between two tuples, $d(x, x')$, used in the k-NN situation testing implementation (k-NN ST) [62]. We do so because we want to compare our implementation, k-NN CST, against its standard counterpart, k-NN ST. We summarize the current algorithmic CST implementation in the Appendix. Let us define the *distance between two tuples* as:

$$d(x, x') = \frac{\sum_{i=1}^{|X|} d_i(x_i - x_i')}{|X|} \tag{6}$$

where (6) averages the sum of the per-attribute distances across X. A lower (6) implies a higher similarity between the tuples x and x'. Here, d_i equals the overlap measurement (ol) if the attribute X_i is categorical; otherwise, it equals the normalized Manhattan distance (md) if the attribute X_i is continuous, ordinal, or interval. We define each $md(x_i, x_{i'}) = |x_i - x_{i'}|/(\max(X) - \min(X))$, and $ol(x_i, x_{i'}) = 1$ if $x_i = x_{i'}$ and 0 otherwise. CST can handle non-normalized attributes but, unless specified, we normalize them to insure comparable per-attribute distances. The choice of (6) is not restrictive. In subsequent works we hope to explore other distance options like , e.g., heterogeneous distance functions [67], as well as probability-based options, e.g., propensity score weighting [51].

CST is, above all, a framework for detecting discrimination. The choice of d as well as ϕ are specific to the implementation of CST. What is important is that the test group is established around the complainant's counterfactual while the control group, like in other discrimination frameworks, is established around the complainant.
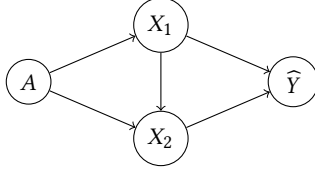
4 EXPERIMENTS

We now showcase the counterfactual situation testing (CST) framework via its k-NN implementation using synthetic (Section 4.1) and real (Section 4.2) datasets. We contrast it to its situation testing counterpart (k-NN ST) [62], and to counterfactual fairness (CF) [39]. Here, for the structural equations $\mathcal{S}$ we assume additive noise. This is a convenient but not necessary assumption that simplifies the abduction step when generating the counterfactuals.[6] We use a significance level of $\alpha = 5\%$, a minimum deviation of $\tau = 0.0$, and a set of k group sizes in $\{15, 30, 50, 100\}$ for CST runs that include and exclude the search centers. Also for comparison, we define individual discrimination as $\Delta p > \tau$ (Def. 3.4) for a single protected attribute. We still, though, demonstrate the use of confidence intervals (Def. 3.5) and how it would affect the final results. Finally, we assume $\mathcal{M}$ and $\mathcal{G}$ in both ADM scenarios.[7]

4.1 An Illustrative Example

We create a synthetic dataset $\mathcal{D}$ based on the scenario in Fig. 1. It is a modified version of Karimi et al. [33, Fig. 1], where we include the protected attribute gender A. Here, gender directly affects both an individual's annual salary X_1 and bank balance X_2, which are used by the bank's ADM $b()$ for approving ($\widehat{Y} = 1$) or rejecting ($\widehat{Y} = 0$) a

---

[5]Formally, $P(\hat{Y}_{A \leftarrow a}(U) = y \mid X, A) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X, A)$, where the left side is the factual $A = a$ and the right side the counterfactual $A = a'$.

[6]Formally, we assume $\mathcal{S} = \{W_j \leftarrow f_j(W_{pa(j)}) + U_j\}_{j=1}^{P}$ for $W = X \cup A$.

[7]The code and data are available at https://github.com/cc-jalvarez/counterfactual-situation-testing.
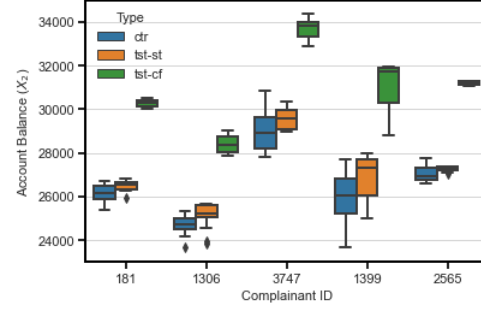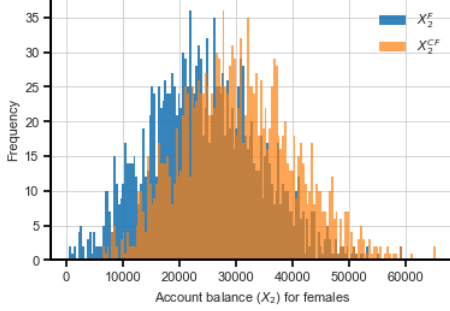
$$\mathcal{M} \begin{cases} A & \leftarrow U_A \\ X_1 & \leftarrow f_1(A) + U_1 \\ X_2 & \leftarrow f_2(X_1, A) + U_2 \end{cases}$$

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

Figure 1: The causal knowledge with corresponding SCM $\mathcal{M}$ and DAG $\mathcal{G}$ behind our (illustrative example) 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 bank's ADM $b()$.

Figure 2: Left. Account balance (X_2) distribution for females in the factual $\mathcal{D}$ and counterfactual $\mathcal{D}^{CF}$ datasets. Right. A comparison on X_2 of the ST and CST (w/o) control group (ctr) versus the ST (tst-st) and CST (w/o) (tst-cf) test groups for five randomly chosen complainants detected by both methods, showing the *fairness given the difference* behind CST as tst-st is closer to ctr than tst-cf.

loan application. We generate $\mathcal{D}$ for $n = 5000$ under $A \sim \text{Ber}(0.45)$ with $A = 1$ if the individual is female and $A = 0$ otherwise, and assume: $X_1 \leftarrow (-\$1500) \cdot \text{Poi}(10) \cdot A + U_1$; $X_2 \leftarrow (-\$300) \cdot \mathcal{X}^2(4) \cdot A + (3/10) \cdot X_1 + U_2$; and $\widehat{Y} = \mathbb{1}\{X_1 + 5 \cdot X_2 > 225000\}$ with $U_1 \sim \$10000 \cdot \text{Poi}(10)$ and $U_2 \sim \$2500 \cdot \mathcal{N}(0, 1)$. Here, $\mathcal{D}$ represents *a known biased scenario.*[8] With A we introduce a systematic bias onto the relevant decision attributes for female applicants.

To run CST we first generate the counterfactual dataset $\mathcal{D}^{CF}$ based on the intervention $do(A := 0)$, or *what would have happened had all loan applicants been male*? Comparing $\mathcal{D}$ to $\mathcal{D}^{CF}$ already highlights the unwanted systematic effects of A. This can be seen, for instance, in Fig. 2 by the rightward shift experienced in X_2 for all female applicants when going from the factual to the counterfactual world. The loan rejection rate for females drops from 60.9% in $\mathcal{D}$ to 38.7% in $\mathcal{D}^{CF}$, which is now closer to the loan rejection rate of 39.2% experienced by males in both worlds. We run CST for all k sizes. Results are shown in Table 1, where w/o refers to "without search centers" for CST.

Does $b()$ discriminate against female applicants? As Table 1 shows, all three methods detect a number of individual discrimination cases. On one hand, the bank clearly uses information that is neutral and needed for approving a loan request; on the other hand, this information is tainted by the effects of gender on such information and the bank, in turn, continues to perpetuate biases

---

[8]Such "penalties", e.g., capture the financial burdens female professionals face in the present after having been discouraged in the past from pursuing high-paying, male-oriented fields [17].

against women in this scenario. The results show how these neutral provisions are harmful toward female applicants, and uncover potential individual discrimination cases.

CST relative to situation testing (ST). Here, consider the CST version without the search centers as ST excludes them. What is clear from Table 1 is that CST finds more individual discrimination cases than ST for all k sizes. For $k = 50$, e.g., CST (w/o) detects 20% while ST just 5%. These results highlight the impact of operationalizing *fairness given the difference*, as the main difference between the two frameworks is how each individual test group is constructed based on the choice of search center. The control group is constructed the same way for both ST and CST.

The choice of the test search centers is what sets CST apart from ST. Note that ST performs an *idealized comparison*. Consider, e.g., the tuple ($x_1 = 35000, x_2 = 7948, a = 1$) as the complainant c. With c as the test search center, the most similar male profiles to the complainant under any distance d would be tuples similar to ($x_1 = 35000, x_2 = 7948, a = 0$). CST, conversely, performs a more *flexible comparison* under *fairness given the difference*. With the corresponding counterfactual tuple ($x_1^{CF} = 50796, x_2^{CF} = 13852, a^{CF} = 0$) as the test search center, the most similar male profiles to the complainant under the same d would be tuples similar to the counterfactual itself. Hence, the test group we construct under CST will probably not be similar to the one we construct under ST nor to the same control group we construct for both ST and CST. Fig. 2, for instance, shows clearly this result for $k = 15$. We randomly chose five complainants that were discriminated by $b()$ according to both ST and CST and plot the distribution of X_2 for the control group

Table 1: Number (and %) of detected individual discrimination cases for the illustrative example based on gender.

Method	$k = 0$	$k = 15$	$k = 30$	$k = 50$	$k = 100$
CST (w/o)	0	288 (16.8%)	313 (18.3%)	342 (20%)	395 (23.1%)
ST [62]	0	55 (3.2%)	65 (3.8%)	84 (5%)	107 (6.3%)
CST	0	420 (24.5%)	434 (25.4%)	453 (26.5%)	480 (28%)
CF [39]	376 (22%)	376 (22%)	376 (22%)	376 (22%)	376 (22%)

Table 2: Subset of individual discrimination cases detected by both CST ($k = 15$) and CF with CI under $\alpha = 5\%$

Comp. (ID)	p_c	p_t	Δp	CI ($\alpha = 5\%$)
44	1.00	0.00	1.00*	[1.00, 1.00]
55	0.81	0.00	0.81*	[0.65, 0.97]
150	1.00	0.94	0.06	[-0.04, 0.16]
203	1.00	0.88	0.13	[-0.01, 0.26]
218	0.56	0.00	0.56*	[0.36, 0.77]

Table 3: Subset of individual discrimination cases detected by CST ($k = 15$) but not by CF with CI under $\alpha = 5\%$.

Comp. (ID)	p_c	p_t	Δp	CI ($\alpha = 5\%$)
5	0.06	0.0	0.06	[-0.04, 0.16]
147	0.50	0.0	0.5*	[0.29, 0.71]
435	0.38	0.0	0.38*	[0.18, 0.58]
1958	0.13	0.0	0.13	[-0.01, 0.26]
2926	0.75	0.0	0.75*	[0.57, 0.93]

(ctr), the ST test group (tst-st), and the CST test group (tst-cf). In this scenario, all 55 ST cases are also detected by CST.

CST relative to counterfactual fairness (CF). Here, consider the CST version including the search centers (though CST w/o is of interest also), as these represent the instances used by CF. We define CF discrimination as a case where the factual $\widehat{y}_c = 0$ becomes $\widehat{y}_c^{CF} = 1$ after the intervention of A. Under this definition, we detect 376 cases of CF discrimination in $\mathcal{D}$, or 22% of female applicants. CF is independent from k as the framework applies only to the individual comparison of the factual and counterfactual tuples for complainant c. Table 1 show that CST detects a higher number of individual discrimination cases for each k size (while CST w/o only passes CF at $k = 100$). In fact, in this scenario, all cases detected by CF are contained in CST.

What sets CST apart from CF is twofold. First, CST equips the CF comparison with certainty measures. This point is illustrated in Table 2 where we show individual cases of discrimination detected by both CF and CST along with confidence intervals (CI) (5) provided by the CST framework. Second, CST detects cases of individual discrimination that are counterfactually fair. This point is illustrated in Table 3 where we show individual cases that pass CF but still exhibit a discriminatory pattern when looking at Δp. Such results highlight why legal stakeholder require multiple comparisons to insure that c's experience is representative of the discrimination claim.

Confidence in results. Finally, notice that Tables 2 and 3 include cases where $\tau = 0$ falls within the individual CI. We detected these

cases under $\Delta p > \tau$ (Def. 3.4). Under $\alpha = 5\%$, we would reject these cases as individual discrimination claims with confidence level of 95% since the minimum deviation is covered by the CIs (Def. 3.5). These are cases with small Δp's that are too close to call. We denote those statistically significant cases with the asterisk on Δp.
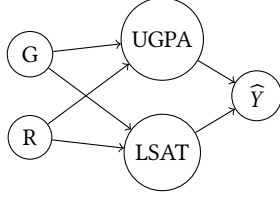
4.2 Law School Admissions

Based on the Law School Success example popularized by Kusner et al. [39, Fig. 2] using US data from the Law School Admission Council survey [66], we create an admissions scenario to a top law school. We consider as protected attributes an applicant's gender, male/female, (G) and race, white/non-white, (R). We add an ADM $b()$ that considers the applicant's undergraduate grade-point average (*UGPA*) and law school admissions test scores (LSAT) for admission. If an applicant is successful, $\widehat{Y} = 1$; otherwise $\widehat{Y} = 0$. We summarize the scenario in Fig. 3. For $b()$, we use the median entry requirements for the top US law school to derive the cutoff ψ.[9] The cutoff is the weighted sum of 60% 3.93 over 4.00 in *UGPA* and 40% 46.1 over 48 *LSAT*, giving a total of 20.8; the maximum possible score under $b()$ is 22 for an applicant. The structural equations follow (1), as in [39], with b_U and b_L denoting the intercepts; $\beta_1, \beta_2, \lambda_1, \lambda_2$ the weights; and $UGPA \sim \mathcal{N}$ and $LSAT \sim$ Poi. the probability distributions.

The dataset $\mathcal{D}$ contains $n = 21790$ applicants, 43.8% are females and 16.1% are non-whites. Despite the ADM $b()$ being externally imposed by us for the purpose of illustrating the CST framework, under $b()$ only 1.88% of the female applicants are successful compared to 2.65% of the male applicants; similarly, only 0.94% of the non-white applicants are successful compared to 2.58% of the white applicants. Therefore, is $b()$ discriminatory toward non-white and female applicants? We run CST along with ST and CF for each protected attribute. We generate the counterfactual dataset $\mathcal{D}^{CF}$ for each R–*what would have been the outcome had all applicants been white?*–and G–*what would have been the outcome had all applicants been male?*–and present the results, respectively, in Table 4 and Table 5. Both tables show CST detecting more individual cases of discrimination than ST and CF.

A similar analysis of Tables 4 and 5 for CST (w/o) versus ST and CST versus CF as in Section 4.1 follows. What the results here highlight, though, is how the two versions of CST compare to each other. In both tables, as k increases, CST (w/o) catches on to CST in the number of cases. This is likely related to how the observations for female/male and non-white/white are distributed in the dataset; though, it also relates to the fact that the difference in size between the groups in each version is just one instance: k versus $k + 1$.

---

[9]That being Yale University Law School: https://www.ilrg.com/rankings/law/index/1/asc/Accept

$$\mathcal{M} \begin{cases} R & \leftarrow U_R \\ G & \leftarrow U_G \\ UGPA & \leftarrow b_U + \beta_1 \cdot R + \lambda_1 \cdot G + U_1, \\ LSAT & \leftarrow \exp\{b_L + \beta_2 \cdot R + \lambda_2 \cdot G + U_2\}, \end{cases}$$

$$\widehat{Y} = \mathbb{1}\{(0.6 \cdot UGPA + 0.4 \cdot LSAT) > \psi\}$$

Figure 3: The causal knowledge with corresponding SCM $\mathcal{M}$ and DAG $\mathcal{G}$ behind the law school admissions dataset, with R denoting race ($R = 1$ for non-white) and G denoting gender ($G = 1$ for female).

Table 4: Number (and %) of individual discrimination cases for the law school admissions scenario based on race.

Method	$k = 0$	$k = 15$	$k = 30$	$k = 50$	$k = 100$
CST (w/o)	0	256 (7.3%)	309 (8.81%)	337 (9.61%)	400 (11.41%)
ST [62]	0	33 (0.94%)	51 (1.45%)	61 (1.74%)	64 (1.83%)
CST	0	286 (8.16%)	309 (8.81%)	337 (9.61%)	400 (11.41%)
CF [39]	231 (6.59%)	231 (6.59%)	231 (6.59%)	231 (6.59%)	231 (6.59%)

Table 5: Number (and %) of individual discrimination cases in for the law school admissions scenario based on gender.

Method	$k = 0$	$k = 15$	$k = 30$	$k = 50$	$k = 100$
CST (w/o)	0	78 (0.82%)	120 (1.26%)	253 (2.65%)	296 (3.10%)
ST [62]	0	77 (0.81%)	101 (1.06%)	229 (2.4%)	258 (2.71%)
CST	0	99 (1.04%)	129 (1.35%)	267 (2.80%)	296 (3.10%)
CF [39]	56 (0.59%)	56 (0.59%)	56 (0.59%)	56 (0.59%)	56 (0.59%)

We should observe the same trend in Table 1 if we continued to increase k. These results show that the different runs of CST can eventually reach the same conclusions under a certain k size. In practice, it means that we could implement one of the two versions of CST without compromising the number of detected individual discrimination cases, though further research is needed.

5 CONCLUSION

We presented counterfactual situation testing (CST), a new framework for detecting individual discrimination in a dataset of classifier decisions. Compared to other methods, CST uncovers more cases even when the classifier is counterfactually fair. It also equips counterfactual fairness with uncertainty measures. CST acknowledges the pervasive effects of the protected attribute by comparing individual instances in the dataset that are observably different in the factual world but hypothetically similar in the counterfactual world. Thus, the results are not too surprising as CST operationalizes *fairness given the difference*, which is a more flexible take on similarity between individuals for testing discrimination than the standard, idealized comparison of two individuals that only differ on their protected status.

Implementation. CST is, above all, a framework for detecting discrimination that advocates for building the test group on the generated counterfactual of the complainant. How similarity is defined, e.g., obviously conditions the implementation. We presented a k-NN version with d as (6); other implementations are possible and still loyal to CST as long as the construction of the control and test groups follows *fairness given the difference*.

Similarly, detecting discrimination is a difficult, context-specific task. That is why for CST we emphasized the role of the expert in constructing the causal graph necessary for generating the counterfactual instances. Indeed, this step could be optimized using, e.g., causal discovery methods [49], but proving discrimination is time consuming and should remain as such given its sensitive role in our society. Further, we are aware that, e.g., the experimental setting could be pushed further by considering higher dimensions or more complex causal structures. What is the point in doing so, though, if that is not the case with current ADM tools being deployed and audited in real life like the recent Dutch scandal [31]? Proving discrimination is not a problem exclusive to (causal) modeling. With CST we wanted to create a framework aware of the multiple angles to the problem of proving discrimination. The cases we have tackled here are intended to showcase what is possible implementation-wise.

Limitations. As future work, promising directions include extending the framework to cases where causal sufficiency does not hold, which is a common risk, and to cases where the decision maker $b()$ is non-binary or of a specific type (e.g., a decision tree). Here, we have also focused on tabular data. Future work should push CST further into more complex datasets to explore the scalability and robustness of the framework.

Research Ethics. In this work we used anonymized benchmark (and synthetic) data designed for public use, complying with applicable legal and ethical rules. We disclosed all details of our method in line with the transparency mandates of the ACM Code of Ethics.

REFERENCES

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

[2] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2018. Automated Test Generation to Detect Individual Discrimination in AI Models. *CoRR* abs/1809.03260 (2018).

[3] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics.* Princeton University Press.

[4] Julia Angwin, Jef Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (2016).

[5] Susan Athey and Guido W Imbens. 2019. Machine learning methods that economists should know about. *Annual Review of Economics* 11 (2019), 685–725.

[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104, 3 (2016), 671–732.

[7] Marc Bendick. 2007. Situation Testing for Employment Discrimination in the United States of America. *Horizons stratégiques* 3, 5 (2007), 17–39.

[8] Marianne Bertrand and Esther Duflo. 2017. Field Experiments on Discrimination. *Handbook of Economic Field Experiments* 1 (2017), 309–393.

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

[10] Emily Black, Samuel Yeom, and Matt Fredrikson. 2020. FlipTest: fairness testing via optimal transport. In *FAT**. ACM, 111–121.

[11] Eduardo Bonilla-Silva. 1997. Rethinking Racism: Toward a structural interpretation. *American Sociological Review* (1997), 465–480.

[12] Alycia N. Carey and Xintao Wu. 2022. The Causal Fairness Field Guide: Perspectives From Social and Formal Sciences. *Frontiers Big Data* 5 (2022), 892837.

[13] Silvia Chiappa. 2019. Path-Specific Counterfactual Fairness. In *AAAI*. AAAI Press, 7801–7808.

[14] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. 2020. Fair regression with Wasserstein barycenters. In *NeurIPS*.

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

[16] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences.* Routledge.

[17] Caroline Criado-Perez. 2019. *Invisible Women.* Vintage.

[18] Alexander D'Amour. 2019. On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives. *CoRR* abs/1902.10286 (2019).

[19] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018).

[20] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *ITCS*. ACM, 214–226.

[21] European Commission. 2021. The Artificial Intelligence Act. https://artificialintelligenceact.eu/. Accessed on January 2nd, 2023.

[22] Michael Fix and Raymond J. Struyk. 1993. *Clear and Convincing Evidence: Measurement of Discrimination in America.* Urban Institute Press.

[23] European Union Agency for Fundamental Rights and Council of Europe. 2018. Handbook on European non-discrimination law. https://fra.europa.eu. Downloaded in 2023..

[24] S. R. Foster. 2004. Causation in antidiscrimination law: Beyond intent versus impact. *Houston Law Review* 41, 5 (2004), 1469–1548.

[25] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *ESEC/SIGSOFT FSE*. ACM, 498–510.

[26] Claudia Goldin and Cecilia Rouse. 2000. Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review* 90, 4 (September 2000), 715–741.

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

[28] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *FAT**. ACM, 501–512.

[29] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction.* Vol. 2. Springer.

[30] James J Heckman. 1998. Detecting Discrimination. *Journal of Economic Perspectives* 12, 2 (1998), 101–116.

[31] Melissa Heikkila. 2022. Dutch scandal serves as a warning for Europe over risks of using algorithms. *POLITICO* (2022).

[32] Lily Hu and Issa Kohler-Hausmann. 2020. What's sex got to do with machine learning?. In *FAT**. ACM, 513.

[33] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: from Counterfactual Explanations to Interventions. In *FAccT*. ACM, 353–362.

[34] Atoosa Kasirzadeh and Andrew Smart. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In *FAccT*. ACM, 228–236.

[35] Niki Kilbertus, Philip J. Ball, Matt J. Kusner, Adrian Weller, and Ricardo Silva. 2019. The Sensitivity of Counterfactual Fairness to Unmeasured Confounding. In *UAI (Proceedings of Machine Learning Research, Vol. 115)*. AUAI Press, 616–626.

[36] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *NIPS*. 656–666.

[37] Jon M. Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2019. Discrimination in the Age of Algorithms. *CoRR* abs/1902.03731 (2019).

[38] Issa Kohler-Hausmann. 2018. Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination. *Nw. UL Rev.* 113 (2018), 1163.

[39] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *NIPS*. 4066–4076.

[40] Christos Louizos, Uri Shalit, Joris M. Mooij, David A. Sontag, Richard S. Zemel, and Max Welling. 2017. Causal Effect Inference with Deep Latent-Variable Models. In *NIPS*. 6446–6456.

[41] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. Survey on Causal-based Machine Learning Fairness Notions. *CoRR* abs/2010.09553 (2020).

[42] Ron Mallon. 2007. A field guide to social construction. *Philosophy Compass* 2, 1 (2007), 93–108.

[43] Lawrence C McCandless, Paul Gustafson, and Adrian Levy. 2007. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine* 26, 11 (2007), 2331–2347.

[44] Deirdre Mulligan. 2022. Invited Talk: Fairness and Privacy. https://www.afciworkshop.org/afcp2022. At the NeurIPS 2022 Workshop on Algorithmic Fairness through the Lens of Causality and Privacy..

[45] Thomas B Nachbar. 2021. Algorithmic fairness, algorithmic discrimination. *Florida State University Law Review* 48 (2021), 50.

[46] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

[47] Judea Pearl, Madelyn Glymour, and Noicholas P. Jewell. 2016. *Causal Inference in Statistics: A Primer.* John Wiley & Sons.

[48] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *KDD*. ACM, 560–568.

[49] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms.* The MIT Press.

[50] Drago Plecko and Elias Bareinboim. 2022. Causal Fairness Analysis. *CoRR* abs/2207.11385 (2022).

[51] Bilal Qureshi, Faisal Kamiran, Asim Karim, Salvatore Ruggieri, and Dino Pedreschi. 2020. Causal inference for social discrimination reasoning. *J. Intell. Inf. Syst.* 54, 2 (2020), 425–437.

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

[53] Dan-Olof Rooth. 2021. Correspondence testing studies. *IZA World of Labor* 58 (2021).

[54] Isabelle Rorive. 2009. Proving Discrimination Cases: The Role of Situation Testing. *Centre for Equal Rights and MPG* (2009). https://ec.europa.eu/migrant-integration/library-document/proving-discrimination-cases-role-situation-testing_en

[55] Evan K. Rose. 2022. A Constructivist Perspective on Empirical Discrimination Research. *Working Manuscript* (2022). https://ekrose.github.io/files/constructivism.pdf

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

[57] Salvatore Ruggieri, José M. Álvarez, Andrea Pugnana, Laura State, and Franco Turini. 2023. Can We Trust Fair-AI?. In *AAAI*. AAAI Press, 15421–15430.

[58] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2010. Data mining for discrimination discovery. *ACM Trans. Knowl. Discov. Data* 4, 2 (2010), 9:1–9:40.

[59] Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. 2017. When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. In *NIPS*. 6414–6423.

[60] Eric C Schneider. 2008. *Smack: Heroin and the American city.* University of Pennsylvania Press.

[61] Maya Sen and Omar Wasow. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science* 19, 1 (2016), 499–522.

[62] Binh Luong Thanh, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *KDD*. ACM, 502–510.

[63] Michael Carl Tschantz. 2022. What is Proxy Discrimination?. In *FAccT*. ACM, 1993–2003.

[64] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[65] White House. 2022. Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/. Accessed on January 2nd, 2023.

[66] Linda F Wightman. 1998. *LSAC national longitudinal bar passage study*. Law School Admission Council.

[67] D. Randall Wilson and Tony R. Martinez. 1997. Improved Heterogeneous Distance Functions. *J. Artif. Intell. Res.* 6 (1997), 1–34.

[68] Raphaële Xenidis. 2020. Tuning EU equality law to algorithmic discrimination: Three pathways to resilience. *Maastricht Journal of European and Comparative Law* 27, 6 (2020), 736–758.

[69] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. 2021. Causal Intersectionality and Fair Ranking. In *FORC (LIPIcs, Vol. 192)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 7:1–7:20.

[70] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation Testing-Based Discrimination Discovery: A Causal Inference Approach. In *IJCAI*. IJCAI/AAAI Press, 2718–2724.