Journal of Artificial Intelligence Research 82 (2025) 2279-2323 Submitted 12/2024; published 04/2025

Counterfactual Situation Testing: From Single to Multidimensional Discrimination

Jose M. Alvarez JOSEMANUEL.ALVAREZ@KULEUVEN.BE
Department of Computer Science, KU Leuven
3001 Leuven, Belgium

Salvatore Ruggieri SALVATORE.RUGGIERI@UNIPI.IT
Department of Computer Science, University of Pisa
56126 Pisa, Italy

Abstract

As machine learning models enable decisions once performed only by humans, it is central to develop tools that assess the fairness of such models. Notably, within high-stake settings like hiring and lending, these tools must be able to detect potentially discriminatory models. 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?" It extends the legally-grounded situation testing (ST) of Thanh et al. (2011) by operationalizing the notion of *fairness given the difference* via counterfactual reasoning. ST finds for each complainant similar protected and non-protected instances in the dataset; constructs, respectively, a control and test group; and compares the groups such that a difference in model outcomes implies a potential case of individual discrimination. CST, instead, avoids this idealized comparison by establishing the test group on the complainant's generated counterfactual, which reflects how the protected attribute when changed influences other seemingly neutral attributes of the complainant. Under CST we test for discrimination for each complainant by comparing similar individuals within the control and test group but dissimilar individuals across these groups. We consider single (e.g., gender) and multidimensional (e.g., gender and race) discrimination testing. For multidimensional discrimination we study multiple and intersectional discrimination and, as feared by legal scholars, find evidence that the former fails to account for the latter kind. Using a k-nearest neighbor implementation, we showcase CST on synthetic and real data. Experimental results show that CST uncovers a higher number of cases than ST, even when the model is counterfactually fair. CST, in fact, extends counterfactual fairness (CF) of Kusner et al. (2017) by equipping CF with confidence intervals, which we report for all experiments.

1. Introduction

Many decisions today are increasingly enabled by machine learning (ML) models. Algorithmic decision-making (ADM) is becoming ubiquitous and its societal discontents clearer (Angwin et al., 2016; Dastin, 2018; Heikkila, 2022). There is a shared urgency by regulators and researchers alike to develop frameworks that asses these ML models for potential discrimination based on protected attributes such as gender, race, and religion (Álvarez et al., 2024; Kleinberg et al., 2019; Ruggieri et al., 2023). Discrimination is often conceived as a causal claim on the effect of the protected attribute over an individual decision outcome

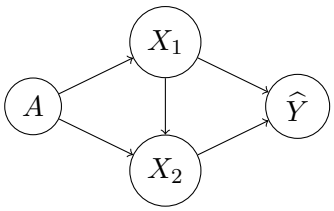
(Heckman, 1998). It is, in particular, a conception based on counterfactual reasoning—what would have been the ML model outcome if the individual, or *complainant*, were of a different protected status?—where we "manipulate" the protected attribute of the individual. Kohler-Hausmann (2018) calls such conceptualization of discrimination the *counterfactual causal model of discrimination* (CMD).

Most frameworks for proving discrimination are based on the CMD. 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 two groups to detect the effects of the protected attribute. Among the available frameworks (Carey & Wu, 2022; Makhlouf et al., 2024; Romei & Ruggieri, 2014), however, there is a need for one that is both *actionable* and *meaningful*. A framework is actionable if it rules out random circumstances from the discrimination claim as required by courts (e.g., EU-FRA (2018), Foster (2004), and Nachbar (2021)) and meaningful if it accounts for known links between the protected attribute and all other attributes as demanded by social scientists (e.g., Bonilla-Silva (1997), Kasirzadeh and Smart (2021), and Sen and Wasow (2016)). In practice, we view actionability as an inferential concern to be handled by comparing multiple control-test instances around a complainant, while meaningfulness as an ontological concern to be handled by requiring domain-knowledge on the protected attribute and its effect on the other attributes of the complainant.

In this work, we present *counterfactual situation testing* (CST), a causal data mining framework for detecting individual cases of discrimination in a dataset of classifier decisions. The dataset can be the one used to train a ML model or the one of actual decisions by a ML model. In the former case we want to prevent learning discriminatory patterns, while in the latter case we want to detect discriminatory decisions. The goal of CST is to be both an actionable and meaningful framework. It combines (structural)[1] counterfactuals (Pearl, 2009) with situation testing (Thanh et al., 2011). *Counterfactuals* answer to counterfactual queries, such as the one motivating the CMD, 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 (Bendick, 2007; Rorive, 2009). For each complainant, given a search algorithm and distance function for measuring similarity, situation testing 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. Hence, *CST follows the situation testing pipeline with the important exception that it constructs the test group around the complainant's counterfactual instead of the complainant.* To illustrate CST and how it compares to standard situation testing, let us use Example 1.1 below.

Example 1.1. (An illustrative example) Let us consider the scenario in Figure 1 in which a bank uses the classifier b to accept or reject ($\hat{Y}$) individual loan applications based on annual salary (X_1) and account balance (X_2), such that $b(X_1, X_2) = \hat{Y}$. Suppose a female applicant ($A = 1$) with $x_1 = 35000$ and $x_2 = 7048$ is rejected and files for gender discrimination. According to Figure 1, the bank uses non-sensitive information to calculate $\hat{Y}$, but

1. Not to be confused with counterfactual explanations (Wachter et al., 2017). Karimi et al. (2021), e.g., use "structural" to differentiate counterfactuals from counterfactual explanations.



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

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

Figure 1: The auxiliary causal knowledge for Example 1.1 (and Section 5.2). Let A denote gender, X_1 annual salary, X_2 account balance, and $\widehat{Y}$ the loan decision by $b()$. It consists of a causal graph (left) and a set of structural equations (right), both introduced in Section 3.

there is also a "known" link between A and $\{X_1, X_2\}$ that questions the neutrality of the information. *Under situation testing*, we find a number of females and males with similar characteristics in terms of X_1 and X_2 to the complainant and compare them. Comparing multiple instances allows to check whether the complainant's claim is representative of an unfavorable pattern toward female applicants by the model (i.e., actionability). However, knowing what we know about A and its influence on X_1 and X_2, would it be fair to compare these similar females and males? As argued by works like Kohler-Hausmann (2018), the answer is no as this *idealized comparison* takes for granted the effect of A on X_1 and X_2 by allowing the former to change while expecting the latter two to remain the same despite the known links. *Under counterfactual situation testing*, instead, we generate the complainant's counterfactual using the auxiliary causal knowledge, creating a male applicant with a higher $x_1 = 50796$ and $x_2 = 13852$, and use him to find similar male instances for constructing the test group. The resulting control and test groups have similar X_1 and X_2 within them but 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). We come back to this example in Section 5.

We evaluate the CST framework on synthetic and real ADM datasets. We use a k-nearest neighbor implementation of the framework, k-NN CST, to compare it to its situation testing counterpart, k-NN ST, by Thanh et al. (2011). The k denotes the number of instances for each control and test groups, determining the size of the many-to-many comparison of each complainant in the dataset. Our experiments show that CST detects a higher number of individual discrimination cases across different k sizes. The results illustrate the impact of moving from the idealized comparison of the k-NN ST to the fairness given the difference comparison of the k-NN CST. This last point is important as legal scholars continue to call for an alternative to the idealized comparison (Kohler-Hausmann, 2018). Importantly, we consider single and multidimensional discrimination, meaning, respectively, claims based on one and many protected attributes. While single discrimination testing is commonly studied, multidimensional discrimination testing is largely unexplored and often portrayed as a straightforward extension to single discrimination testing (Xenidis, 2020).

Multidimensional discrimination covers two forms: multiple and intersectional. In multiple discrimination the complainant must be discriminated for each of the protected attributes, while in intersectional discrimination the complainant must be discriminated at

the intersection of the protected attributes. To illustrate this distinction, suppose the complainant in Example 1.1 is also non-white and makes a claim based on gender and race. Multiple discrimination occurs if the complainant is discriminated, separately, as a female and as a non-white individual. Intersectional discrimination occurs if the complainant is discriminated, simultaneously, as a female-non-white individual. Each form of multidimensional discrimination, in turn, poses different problem formulations for discrimination testing (Roy et al., 2023; Wang et al., 2022). Beyond the distinct problem formulations, an open issue with these two forms of discrimination is that only multiple discrimination is recognized by non-discrimination law. Legal scholars have raised concerns on this lack of recognition for intersectional discrimination, arguing that multiple discrimination fails to account for it (Xenidis, 2020). We test for multiple and intersectional discrimination using CST, finding that the former does not capture the latter. This work is the first to evaluate and provide evidence for this legal concern.

Additionally, CST provides an actionable extension to *counterfactual fairness* by Kusner et al. (2017), which remains the leading causal fair ML definition (Makhlouf et al., 2024). A ML model is counterfactually fair when the complainant's and its counterfactual's decision outcomes are the same. These are the same instances used by CST to construct, respectively, the control and test groups, which allows to equip this popular fairness definition with measures for uncertainty due to the many-to-many comparison. CST links counterfactual fairness claims with statistical significance, and positions it for discrimination testing as uncertainty measures are often required by courts (EU-FRA, 2018). By looking at the control and test groups rather than the literal comparison of the factual versus counterfactual instances, CST evaluates whether the counterfactual fairness claim itself is representative of similar instances. Our results show that individual discrimination can occur even when the ML model is counterfactually fair, capturing the scenario where a model discriminates when evaluating borderline instances.

In summary, with CST we present a meaningful and actionable framework for detecting individual discrimination. Our main contributions are threefold. First, we offer the first explicit operationalization of *fairness given the difference* for discrimination testing and, in doing so, define a new view on similarity that is more flexible than the standard idealized comparison. Second, we explore single and multidimensional discrimination testing, studying the latter's tension between multiple and intersectional discrimination. Third, we equip counterfactual fairness with confidence intervals, introducing an actionable extension to the popular causal fairness definition.

The rest of the paper is organized as follows. We present the related work in Section 2 and the role of auxiliary causal knowledge within CST for discrimination testing in Section 3. We introduce the CST framework in Section 4, including its k-NN implementation. We showcase CST using two classification scenarios in Section 5. We discuss the main limitations of this work in Section 6. We conclude in Section 7.

2. Related Work

We position CST with respect to current frameworks for discrimination testing along the goals of actionability and meaningfulness. Later in Section 3 we discuss the role of causality for conceiving discrimination. For a broader, multidisciplinary view on discrimination

testing, we refer to the survey by Romei and Ruggieri (2014). For a recent survey of the fair ML testing literature, see Chen et al. (2024).

Regarding actionability, it is important when proving discrimination to insure that the framework accounts for sources of randomness in the decision-making process. Popular non-algorithmic frameworks—such as natural (Goldin & Rouse, 2000) and field (Bertrand & Duflo, 2017) experiments, audit (Fix & Struyk, 1993) and correspondence (Bertrand & Mullainathan, 2004; Rooth, 2021) studies—address this issue by using multiple observations to build inferential statistics. Similar statistics are asked in court for proving discrimination (EU-FRA, 2018, Section 6.3). Few algorithmic frameworks address this issue due to model complexity preventing formal inference (Athey & Imbens, 2019). An exception are data mining frameworks for discrimination discovery (Pedreschi et al., 2008; Ruggieri et al., 2010) that operationalize the non-algorithmic notions, including situation testing (Thanh et al., 2011; Zhang et al., 2016). These frameworks (Aggarwal et al., 2018; Galhotra et al., 2017; Qureshi et al., 2020) 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. It remains unclear if the same can be said about existing causal fair machine learning methods as these have yet to be used beyond academic circles. The suitability of algorithmic fairness methods for testing discrimination, be it or not ADM, remains an ongoing discussion (Weerts et al., 2023).

Regarding meaningfulness, situation testing and the other methods previously mentioned have been criticized for their handling of the counterfactual question behind the causal model of discrimination (Hu & Kohler-Hausmann, 2020; Kasirzadeh & Smart, 2021; Kohler-Hausmann, 2018). In particular, these actionable methods take for granted the influence of the protected attribute on all other attributes. This point 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 an approach goes against how most social scientists interpret the protected attribute and its role as a social construct when proving discrimination (Bonilla-Silva, 1997; Hanna et al., 2020; Rose, 2022; Sen & Wasow, 2016). It is in that regard where structural causal models (Pearl, 2009) and their ability to generate counterfactuals via the abduction, action, and prediction steps (e.g., Chiappa (2019) and Yang et al. (2021)), including counterfactual fairness (Kusner et al., 2017), have an advantage. This advantage is overlooked by critics of counterfactual reasoning (Hu & Kohler-Hausmann, 2020; Kasirzadeh & Smart, 2021): generating counterfactuals, as long as the structural causal model is properly specified, accounts for the effects of changing the protected attribute on all other attributes. Hence, a framework like counterfactual fairness, relative to situation testing and other discrimination discovery methods, is more meaningful in its handling of protected attributes.

CST bridges these two lines of work, borrowing the actionability aspects from frameworks like situation testing and meaningful aspects from frameworks like counterfactual fairness. Counterfactual generation allows to create a comparator for the complainant that accounts for the influence of the protected attribute on the other attributes, departing from the idealized comparison. It is not far, conceptually, from the broader ML problem of learning fair representations (Zemel et al., 2013) since we wish to learn (read, map) a new representation of the complainant that reflects where it would have been had it belonged to the non-protected group. It is a normative claim on what a non-protected instance similar

to the complainant looks like. Beyond counterfactual fairness and derivatives (e.g., Chiappa (2019)), other works address this problem of deriving such a pair for the complainant. For instance, Plečko and Meinshausen (2020) use a quantile regression approach while Bothmann et al. (2023) use a residual-based approach for generating the pair. Both works rely on having access to a structural causal model, but do not exploit the abduction, action, and prediction steps for generating counterfactual distributions. Black et al. (2020), instead, propose the FlipTest, a non-causal approach that uses an optimal transport mapping to derive the pair for the complainant. These three works exemplify ML methods that use counterfactual reasoning to operationalize different interpretations of individual similarity. With CST we align with these and similar efforts to propose an alternative to the idealized comparison often used in discrimination testing.

3. Causal Knowledge for Discrimination Testing

In this section, we discuss the role of auxiliary causal knowledge within CST. We formulate causality using structural causal models (SCM). CST requires access to the dataset of decisions, $\mathcal{D}$, and the ML model that produced it, $b()$. CST also requires a SCM describing the data generating model behind $\mathcal{D}$. We view this last requirement as an input space for stakeholders and domain experts. SCM are a convenient way for organizing assumptions on the source of the discrimination, facilitating stakeholder participation and supporting collaborative reasoning about contested concepts (Mulligan, 2022). There is no ground-truth concerning the SCM for $\mathcal{D}$. The SCM describes an agreed view on the discrimination problem, though, not necessarily the only nor correct view of it.

Let $\mathcal{D}$ contain the set of relevant attributes X, the set of protected attributes A, and the decision outcome $\hat{Y}$ such that $\hat{Y} = b(X)$. We describe $\mathcal{D}$ as a collection of n tuples, each $(x_i, a_i, \widehat{y}_i)$ representing the i^{th} 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 gender). We relax this assumption in Section 4.5 when formalizing multidimensional discrimination.

3.1 Structural Causal Models and Counterfactuals

A *structural causal model* (SCM) (Pearl, 2009) $\mathcal{M} = \{\mathcal{S}, \mathcal{P}_{\mathbf{U}}\}$ describes how the set of p variables $W = X \cup A$ is determined based on corresponding sets of structural equations $\mathcal{S}$, and p 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 latent variable U_j with distribution $P(U_j) \in \mathcal{P}_{\mathbf{U}}$. Formally, for $W_j \in W$ we have that:

$$W_j \leftarrow f_j(W_{pa(j)}, U_j) \tag{1}$$

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 can use $\mathcal{M}$ to derive $\mathcal{G}$.[2]

We make three assumptions for the SCM $\mathcal{M}$ that are common within the causal fairness literature (Makhlouf et al., 2024). First, we assume *causal sufficiency*, meaning there are

2. This is based on the global Markov and faithfulness properties, summarized in the notion of the d-separation. We skip d-separation as we do not use it in this paper. See Peters et al. (2017).

no hidden common causes in $\mathcal{M}$, or confounders. Second, we assume $\mathcal{G}$ to be *acyclical*, which turns $\mathcal{G}$ into a directed acyclical graph (DAG), allowing for no feedback loops. Third, we assume *additive noise models* (ANM) to insure an invertible class of SCM (Hoyer et al., 2008). The ANM assumption implies $\mathcal{S} = \{W_j \leftarrow f_j(W_{pa(j)}) + U_j\}_{j=1}^{p}$ in (1). These assumptions are not necessary for generating counterfactuals, but do simplify the process (Pearl et al., 2016). The CST framework is not tied to any of these assumptions as long as the generated counterfactuals are reliable.

The causal sufficiency assumption is particularly deceitful as it is difficult to both test and account for a hidden confounder (D'Amour, 2019; Louizos et al., 2017; McCandless et al., 2007). The risk of a hidden confounder is a general modeling problem. Here, the dataset $\mathcal{D}$ delimits our context. By this we mean that we expect it to contain all relevant information used by $b()$. Causal sufficiency implies independence among the random variables in U, which allows to factorize $P_{\mathbf{U}}$ into its individual components:

$$P(U_1, \ldots, U_j) = P(U_1) \times \cdots \times P(U_j). \tag{2}$$

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 ask such questions around the protected attribute A. By setting A to the non-protected status α using the *do-operator* $do(A := \alpha)$ (Pearl, 2009), we capture the individual-level effects A has on X according to the SCM $\mathcal{M}$. Let X^{CF} denote *the set of counterfactual variables* obtained via the three steps: abduction, action, and prediction (Pearl et al., 2016). Further, let $P(X^{CF}_{A\leftarrow\alpha}(U) \mid X, A)$ denote *counterfactual distribution*. We now describe each step. *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* $P(X^{CF}_{A\leftarrow\alpha}(U) \mid X, A)$ by propagating the abducted $P(U_i \mid X, A)$ through the revised structural equations in $\mathcal{M}'$. Unlike counterfactual explanations (Wachter et al., 2017), generating counterfactuals and, thus, CST, does not require a change in the individual decision outcome.

3.2 On Conceiving Discrimination

Discrimination is a comparative process (Lippert-Rasmussen, 2006). Non-discrimination law is centered on Aristotle's maxim of treating similar (or similarly situated) individuals similarly (Westen, 1982). Granted that we can agree on what similar (or similarly situated) individuals are,[3] in practice, testing for discrimination reduces to comparing similar protected and non-protected by non-discrimination law individuals to see if their outcomes differ within the context of interest (Weerts et al., 2023). Most, if not all, discrimination tools operationalize this comparative process (Kohler-Hausmann, 2018).

The legal setting of interest in this work is indirect discrimination under EU non-discrimination law. Indirect discrimination occurs when an apparently neutral practice disadvantages individuals that belong to a protected group. Following Hacker (2018), we focus on indirect discrimination for three reasons. First, unlike disparate impact under

3. We briefly discuss this issue in the next section, though similarity in itself is a complex, ongoing, legal discussion. We recommend Westen (1982) for further reading.

US law (Barocas & Selbst, 2016), 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 ML models are not allowed to use the protected attribute as input, making it difficult for regulators to use the direct discrimination setting.[4] 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 deriving $\hat{Y}$ based on X. The indirect setting best describes how biased information can still be an issue for a ML model that never uses the protected attribute. That said, it does not mean CST cannot be implemented in other legal contexts. We simply acknowledge that it was developed with the EU non-discrimination legal framework in mind due to the previous reasons.

Causality is often used for formalizing the problem of discrimination testing. This is because of the legal framing of discrimination in which we are interested in the protected attribute as a direct or indirect cause of the decision outcome (Heckman, 1998; Kohler-Hausmann, 2018). Previous works (Chiappa, 2019; Kilbertus et al., 2017; Plecko & Bareinboim, 2024; Tschantz, 2022) focus more on whether the paths between A and $\hat{Y}$ are direct or indirect, leading to the two kinds of discrimination prescribed under EU non-discrimination law. The causal setting here is much simpler. We know that $b()$ only uses X, and are interested in how information from A is carried by X and how we account for these links when testing for discrimination by using the auxiliary causal knowledge.

3.3 Fairness Given the Difference

The SCM required by CST allows to operationalize the notion of *fairness given the difference*, FGD for short, and depart from the standard idealized comparison. Access to a SCM enables the generation of a counterfactual instance for the complainant, allowing to represent how the protected attribute influences the non-protected attributes used by $b()$. We come back to this point in the next section; here, we motivate FGD. The reference work is Kohler-Hausmann (2018). FGD captures that work's overall criticism toward the counterfactual causal model of discrimination (CMD) introduced in Section 1.[5]

As argued by Kohler-Hausmann (2018) and others before her (Bonilla-Silva, 1997; Sen & Wasow, 2016), it is difficult to deny that most protected attributes, if not all of them, are *social constructs*. These are attributes that were used to classify *and* divide groups of people in a systematic way that conditioned the material opportunities of multiple generations (Mallon, 2007; Rose, 2022). 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. Protected attributes summarize the historical processes that fairness researchers are trying to address today and should not be treated lightly.[6]

4. This point, though, has been contested recently by Adams-Prassl et al. (2023).
5. We attribute this phrase to Kohler-Hausmann as she used it during a panel discussion at the NeurIPS'21 Workshop on Algorithmic Fairness through the Lens of Causality and Robustness (AFCR). It is not, however, present in her paper. The phrase first appears in Álvarez and Ruggieri (2023).
6. An example is the use of race by US policy makers after WWII. See, e.g., the historical evidence provided by Rothstein (2017) (for housing), Schneider (2008) (for narcotics), and Adler (2019) (for policing).

FGD centers on how A is treated in the CMD. It goes beyond the standard manipulation concern in which A is an inmutable attribute (Angrist & Pischke, 2008). Instead, granted that we *can* or, more precisely, *have to* manipulate A for running a discrimination analysis, FGD puts into question how a testing framework operationalize such manipulation. If A is a social construct with clear influence on X, then *when A changes, X should change as well.* This is precisely what FGD entails. As discussed in Section 4, within CST it manifests by building 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.

Based on FGD we consider two types of manipulations that summarize existing discrimination testing 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 (Thanh et al., 2011; Zhang et al., 2016) and the famous correspondence study by Bertrand and Mullainathan (2004). 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 (Kusner et al., 2017) uses this manipulation. The MM is clearly preferred over the CP manipulation when we view A as a social construct. See Álvarez and Ruggieri (2024) for a detailed discussion on the CP and MM manipulations.

4. Counterfactual Situation Testing

The goal of CST is to construct and compare a control and test group for each protected individual (read, *complainant*) c in the dataset in a meaningful and actionable way. The focus is on the tuple $(x_c, a_c, \widehat{y}_c) \in \mathcal{D}$, with $c \in [1, n]$, that motivates the individual discrimination claim. CST requires access to the ADM $b()$, the dataset $\mathcal{D}$, and the auxiliary causal knowledge SCM $\mathcal{M}$ and DAG $\mathcal{G}$.

Three additional inputs are central to CST: the number of instances, k; the similarity function, d; and the strength of the evidence for the discrimination claim, α. Here, k determines the size of the control and test groups for c; d determines how much these two groups resemble c; and α determines the statistical significance required when comparing these two groups to trust the claim around c. We must also define a search algorithm for implementing CST. We use the k-nearest neighbors, or k-NN, algorithm (Hastie et al., 2009), resulting in the present k-NN CST. The k-NN is intuitive, easy to implement, and commonly used by other frameworks. The k-NN implementation is straightforward. We provide the relevant algorithms in Appendix B. Other implementations are possible as long as the following definitions are adjusted.

4.1 Measuring Individual Similarity

We start by defining the *similarity measure* d. We use the same d as the one used by Thanh et al. (2011) to compare our implementation against its standard situation testing counterpart. Let us define the *between tuple distance* $d(x_1, x_2)$ as:

$$d(x_1, x_2) = \frac{\sum_{i=1}^{|X|} d_i(x_{1,i}, x_{2,i})}{|X|} \tag{3}$$

such that $d(x_1, x_2)$ averages the sum of the *per-attribute distances* $d_i(x_{1,i}, x_{2,i})$ across all attributes in X. It does not use the protected attribute(s) A. A lower d implies a higher similarity between the tuples x and x' and further implies two similar individuals. The k-NN CST handles non-normalized attributes but, as default, we normalize them to insure comparable per-attribute distances.

The specific d_i used depends on the type of the *i-th* attribute. It 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. Under this conception, d amounts to Gower's distance (Gower, 1971). For illustrative purposes, we recall both *md* and *ol* distances below. We define *md* as:

$$md(x_{1,i}, x_{2,i}) = \frac{|x_{1,i} - x_{2,i}|}{(\max(X_i) - \min(X_i))} \tag{4}$$

and we define *ol* as:

$$ol(x_{1,i}, x_{2,i}) = \begin{cases} 1 & \text{if } x_{1,i} \neq x_{2,i} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The choices of d_i and, in turn, of d are not restrictive. We plan to explore other formulations in subsequent works, like heterogeneous distance functions (Wilson & Martinez, 1997) and propensity score matching (Qureshi et al., 2020). Hence, why we view d as an input to rather than a characteristic of k-NN CST.

4.2 Building the Control and Test Groups

For a complainant c, the control and test groups are built on the search spaces and search centers for each group. The search spaces are derived from $\mathcal{D}$. 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 its corresponding generated counterfactual dataset $\mathcal{D}^{CF}$.

Definition 4.1 (Search Spaces)**.** Under a binary A, with $A = 1$ denoting 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 4.2 (Counterfactual Dataset)**.** Given the ADM $b()$ and SCM $\mathcal{M}$, the *counterfactual dataset* $\mathcal{D}^{CF}$ includes the counterfactual mapping of each instance with a protected status in $\mathcal{D}$ via the abduction, action, and prediction steps when setting a binary A to the non-protected status, or $do(A := 0)$.

Definition 4.3 (Search Centers)**.** 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 counterfactual $(x_c^{CF}, a_c^{CF}, \hat{y}_c^{CF}) \in \mathcal{D}^{CF}$ as the *test search center* for exploring $\mathcal{D}_t \subset \mathcal{D}$.

We extend these definitions for $|A| > 1$ in Section 4.5. Importantly, to obtain $\mathcal{D}^{CF}$ we consider a SCM $\mathcal{M}$ in which A has no causal parents, A affects only elements of X, and $\hat{Y} = b(X)$. See, e.g., Figures 1 and 6. Under such auxiliary causal knowledge, if A changes then X changes too. Here, $\mathcal{D}^{CF}$ represents the world that the complainants would have experienced had they belonged to the non-protected group. This is why we draw the test search center from $\mathcal{D}^{CF}$ and not from $\mathcal{D}$.

Given the $\mathcal{D}$ and $\mathcal{D}^{CF}$, we construct the control and test groups for c using the k-NN algorithm with distance function d (3). We want each group (read, neighborhood) to have 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 the search center to explore $\mathcal{D}_c$:

$$k\text{-}ctr = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D}_c : rank_d(x_c, x_i) \leq k\} \tag{6}$$

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. 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 the search center to explore $\mathcal{D}_t$:

$$k\text{-}tst = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D}_t : rank_d(x_c^{CF}, x_i) \leq k\} \tag{7}$$

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 d for each group. Neither A nor $\hat{Y}$ are used for constructing the groups. Further, we can always expand (6) and (7) by adding constraints such as, for instance, a maximum distance $\epsilon > 0$: $k\text{-}ctr = \{x_i \in \mathcal{D}_c : rank_d(\mathbf{x}_c, x_i) \leq k \wedge d(x_c, x_i) \leq \epsilon\}$ and $k\text{-}tst = \{x_i \in \mathcal{D}_t : rank_d(x_c^{CF}, x_i) \leq k \wedge d(x_c^{CF}, x_i) \leq \epsilon\}$.

Remark 4.1. (Meaningfulness) The choice of search centers for (6) and (7) operationalizes *fairness given the difference* in CST, making it a meaningful framework for testing individual discrimination. Using x_c and x_c^{CF} to search for, respectively, protected and non-protected individuals in $\mathcal{D}_c$ and $\mathcal{D}_t$ is a statement on how we view the role of *within group ordering* as imposed by the protected attribute A. Each search center reflects the A-specific ordering imposed on the search space it targets.

To illustrate Remark 4.1, let us consider Example 1.1. If being a female imposes certain systematic limitations that hinder x_c, then comparing c to other females preserves the group ordering prescribed by $X|A = 1$ as all instances involved experience A in the same way. Similarly, the generated counterfactual male instance for c should reflect the group ordering prescribed by $X|A = 0$. Here, in particular, we would expect $x_c \leq x_c^{CF}$ given what we know about the effects of A on X. A way to reason about this remark is through the notion of effort. If being female requires a higher individual effort than being male 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 undertaken by c to be comparable to these male instances. The counterfactual x_c^{CF} reflects said adjustment for c. For a formal discussion on effort and its role on individual fairness (Dwork et al., 2012), see Chzhen and Schreuder (2022) and Chzhen et al. (2020).

4.3 Detecting Discrimination

For a complainant c, we compare the control and test groups by looking at the *difference in proportion of negative decision outcomes*:

$$\Delta p = p_c - p_t \tag{8}$$

such that p_c and p_t represent the count of tuples with a negative decision outcome, respectively, in the control group (6) and test group (7). Formally:

$$p_c = \frac{|\{(x_i, a_i, \widehat{y}_i) \in k\text{-}ctr : \hat{y}_i = 0\}|}{k} \qquad p_t = \frac{|\{(x_i, a_i, \widehat{y}_i) \in k\text{-}tst : \hat{y}_i = 0\}|}{k} \tag{9}$$

where only $\hat{Y}$ is used for deriving the proportions. It follows that $p_c, p_t \in [0, 1]$ and $\Delta p \in [-1, 1]$. We compute Δp for all complainants in $\mathcal{D}$ regardless of their decision outcome.

CST uses Δp to test for the complainant's individual discrimination claim. Implicit to this task is the *accepted deviation* $\tau \in [-1, 1]$ for Δp. It represents the maximum acceptable difference between p_c and p_t, such that any deviation from it amounts to discrimination: i.e., $\Delta p > \tau$. The τ is often implied with the default choice of $\tau = 0$, as we wish for protected and non-protected individuals to be rejected at equal rates. As Δp is a proportion comparison, Δp is asymptotically normally distributed, which allows to build *Wald confidence intervals* (CI) around it. For other confidence interval methods, such as exact methods for small samples, see Newcombe (1998). With the CI we equip the complainant's claim with a measure of certainty and answer whether the claim, meaning the deviation from τ, is or not statistically significant. If τ falls within the *one-sided CI*, then we cannot say that the complainant's claim is statistically significant. We write such CI for Δp as:

$$[\Delta p - w_\alpha, +\infty), \quad \text{with} \quad w_\alpha = z_\alpha \sqrt{\frac{p_c(1-p_c) + p_t(1-p_t)}{k}}. \tag{10}$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$ is the $1-\alpha$ quantile of the standard normal distribution $\mathcal{N}$ under a *significance level* of α or, equivalently, a *confidence level* $(1-\alpha)\cdot 100\%$.[7] The $+\infty$ represents that there is no upper bound, as we are interested in values greater than τ. The choice of α and τ, as with k, depends on the context of the discrimination claim. These parameters are motivated by legal requirements (set, e.g., by the court (Thanh et al., 2011)), or technical requirements (set, e.g., via power analysis (Cohen, 2013)), or both. A common choice for α is 0.05, though common alternatives are also 0.01 and 0.10.

The CI represents a one-sided statistical test based on the hypothesis that there is individual discrimination, providing a measure of certainty on Δp through a range of possible values. Formally, let π be the true difference in proportion of negative decision outcomes between the control and test groups. Then the *null hypothesis* is $H_0 : \pi = \tau$, while the *alternative hypothesis* $H_1 : \pi > \tau$. When τ falls within the range of probable values in CI, we fail to reject H_0 with α significance level. Given Δp (8) and its CI (10), we can now proceed to define individual discrimination under CST.

Remark 4.2. (Two Versions of CST) CST can include or exclude the search centers in (8) and (10). If we exclude them, then both 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 of $k+1$ in both as well as in the w_α calculation. To distinguish between the two versions of CST, we will use CST w/o when

7. Álvarez and Ruggieri (2023) contains a typo in the numerator of w_α: we wrote a minus instead of a plus sign. In the code, however, it was implemented correctly. It also discusses a two-sided CI.

excluding and CST w/ when including the search centers. We add this option for comparing CST against situation testing (Thanh et al., 2011), which excludes the search centers, and counterfactual fairness (Kusner et al., 2017), which only uses the search centers. We use this option extensively in Section 5.

Definition 4.4 (Individual Discrimination)**.** There is (potential) individual discrimination toward the complainant c if $\Delta p > \tau$, meaning the negative decision outcomes rate for the control group is greater than for the test group by some accepted deviation $\tau \in [-1, 1]$.

Definition 4.5 (Confidence on the Individual Discrimination Claim)**.** A detected (potential) discrimination claim for the complainant c by Definition 4.4 is statistically significant with significance level α if the CI excludes τ.

We highlight the use of the word *potential* in both definitions. It implies, formally, that under CST, as with any individual or group discrimination testing framework (Romei & Ruggieri, 2014), we test for *prima facie* discrimination. Uncovering discrimination amounts to a series of steps among which there is the need to provide evidence of the discrimination claim. Even if said evidence is found, it still needs to be argued for in court. For a discussion on the EU discrimination testing pipeline, see Weerts et al. (2023).

Remark 4.3. (Actionability) The many-to-many comparison behind Δp is what makes CST an actionable framework for testing individual discrimination. The single comparison is not enough when proving *prima facie* discrimination as 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 of the complainant on which we are confident enough.

The notion of repetition is important in Remark 4.3. Similar to flipping a coin multiple times to uncover its (un)fairness, we expect a significant pattern of unfavorable decisions (read, discrimination) to emerge through "repeating" the decision-making process in question. Such repetition is often not possible in practice. In a non-ADM setting we cannot ask the same female complainant in Example 1.1 to apply multiple times to the same bank; we can, instead, look at other similar instances that underwent the same decision process. Similarly, even when repetition is deterministic, such as entering the same input multiple times into the ADM, it is non-trivial to generalize that the individual case represents a group-wise pattern. What rules out that the Δp for complainant c is an exception rather than a systematic effect? We can, for instance, 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. The p_c and p_t help tackle these concerns.

On positive discrimination. Positive individual discrimination, or affirmative action, refers to the setting in which complainant c is shown to be favored in the decision-making process (Romei & Ruggieri, 2014). Policies like diversity quotas are an example of positive discrimination. We can operationalize positive discrimination easily under the current k-NN CST implementation: we would rewrite Definitions 4.4 and 4.5 by looking at the same complainant c but focusing on the opposite effect. Formally, we would consider $\Delta p < \tau$ (where now $\tau < 0$) and, in turn, the CI $(-\infty, \Delta p + w_\alpha]$ as we would test for the alternative hypothesis $H_1 : \pi < \tau$. The rest of the k-NN CST pipeline would apply the same.

Positive discrimination remains understudied within algorithmic discrimination. We believe this is due to standard discrimination being more prevalent as a societal and research problem. We also believe positive discrimination poses a different set of conceptual challenges over traditional discrimination, making it harder to justify by those in favor of it.[8] This is, at least, our reading from the lack of discussion positive discrimination enjoys by the legal works we cite. In short, although testing for positive discrimination under CST is straightforward, a clear legal narrative is lacking for us to comfortably operationalize it. We formalize and test for positive discrimination in Appendices B and C, respectively. We do so mainly for illustrative purposes since we are focused on understanding traditional discrimination in its single and multidimensional forms.

4.4 Connection to Counterfactual Fairness

There is a clear link between CST and counterfactual fairness (CF) of Kusner et al. (2017). Recall that the ADM $b()$ is counterfactually fair if it outputs the same outcome for the factual tuple as for its counterfactual tuple, where the latter is generated through the abduction, action, and prediction steps when intervening the protected attribute A.[9] Hence, the factual $(x_c, a_c, \hat{y}_c)$ and counterfactual $(x_c^{CF}, a_c^{CF}, \hat{y}_c^{CF})$ tuples used in CST are also the ones used in CF for evaluating the counterfactual fairness for complainant c.

We view CST, when including the search centers, as an actionable extension of CF. CST equips CF with CI (10), providing a certainty measure on the counterfactual fairness of $b()$. Previous works on CF have addressed uncertainty concerns (Kilbertus et al., 2019; Russell et al., 2017), but with a focus on the structure of the SCM $\mathcal{M}$ and how that affects the measured unfairness of $b()$. We instead address certainty on the literal comparison that motivates the CF definition. A key consequence of this link between CST and CF is that we can have an ADM $b()$ that is counterfactually fair but discriminatory. We summarize this point in Proposition 4.1. We also present a sketch of proof for it.

Proposition 4.1 (Actionable Counterfactual Fairness)**.** Counterfactual fairness does not imply nor it is implied by individual discrimination as conceived in Definition 4.4.

We now present a sketch of proof to Proposition 4.1. Consider the factual tuple $(x_c, a_c = 1, \widehat{y}_c = 0)$ and assume the generated counterfactual is $(x_c^{CF}, a_c^{CF} = 0, \widehat{y}_c^{CF} = 0)$. Since $\widehat{y}_c = \widehat{y}_c^{CF}$, this is a case in which CF holds. However, the decision boundary of the ADM $b()$ can be purposely set such that the k-nearest neighbors of x_c are all within the decision $\hat{Y} = 0$, and less than $1 - \tau$ fraction of the k-nearest neighbors of x_c^{CF} are within the decision $\hat{Y} = 0$. This leads to a $\Delta p > 1 - (1 - \tau) = \tau$, showing that there is individual discrimination. Similarly, the other way can be shown by assuming $\widehat{y}_c \neq \widehat{y}_c^{CF}$ but the sets of k-nearest neighbors have rates of negative decisions whose difference is smaller than τ.

Proposition 4.1 alludes to the scenario in which $b()$ is counterfactually fair yet discriminatory. 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 closer to the decision

8. Take, e.g., the US Supreme Court's overturn of affirmative action (Totenberg, 2023).
9. Formally, $P(\hat{Y}_{A\leftarrow a}(U) = y \,|\, X, A) = P(\hat{Y}_{A\leftarrow a'}(U) = y \,|\, X, A)$, where the left side is the factual $A = a$ and the right side the counterfactual $A = a'$. Similar with τ in Δp, the equality can be relaxed given some permissible difference threshold $\epsilon > 0$ between the factual and counterfactual quantities.

boundary than the former. 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.

Importantly, for the purposes of Proposition 4.1 we consider the *two-sided CI* over the previous one-sided CI (10). We are interested in addressing the statistical significance of the "counterfactual fairness claim" using the neighborhoods built by the k-NN—note via the CST w/ version, cfr. Remark 4.2—algorithm around the factual and counterfactual instances of CF. We write such CI for Δp as:

$$[\Delta p - w_{\alpha/2}, \Delta p + w_{\alpha/2}], \quad \text{with} \quad w_{\alpha/2} = z_{\alpha/2}\sqrt{\frac{p_c(1-p_c) + p_t(1-p_t)}{k+1}} \tag{11}$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ be the $1-\alpha/2$ quantile of $\mathcal{N}$ under a *significance level* of α or, equivalently, a *confidence level* $(1-\alpha)\cdot 100\%$. We use (11) when addressing CF as a whole, and use (10) when addressing the discrimination claim through CF.

4.5 The Multidimensional Setting

Let us revisit the previous definitions under multidimensional discrimination. It occurs when $|A| > 1$. Following the legal literature (Xenidis, 2020), we distinguish two forms of multidimensional discrimination: multiple and intersectional.

Definition 4.6 (Multiple Discrimination)**.** Under Definition 4.4, there is (potential) multiple individual discrimination toward the complaint c with the set of $|A| = q > 1$ protected attributes, if $\Delta p > \tau$ for each $\{A_i\}_{i=1}^{q}$ protected attribute.

Definition 4.7 (Intersectional Discrimination)**.** Under Definition 4.4, there is (potential) intersectional individual discrimination toward the complaint c with the set of $|A| = q > 1$ protected attributes, if $\Delta p > \tau$ for the attribute $A^* = \mathbb{1}\{A_1 = 1 \wedge A_2 = 1 \wedge \cdots \wedge A_q = 1\}$ obtained by the intersection of the protected attributes.

Regarding Definition 4.5, meaning the confidence on the individual claim under multiple and intersectional discrimination, notice that both Definitions 4.6 and 4.7 work with Δp point estimates. Definition 4.6 looks at q deltas for c given the q protected attributes, while Definition 4.7 looks at a single delta for the attribute obtained by the intersection for c of all the protected attributes. For intersectional discrimination, the single intersection delta must be statistically significant given α. For multiple discrimination, instead, all the multiple deltas must be statistically significant given α/q. The *Bonferroni correction factor* $1/q$ counteracts the well-known multiple comparisons problem, allowing to test for a family-wise error rate of α in a set of q (possibly, dependent) statistical tests. We assume the same τ in both cases, in particular, with τ being the same irrespective of the protected attribute considered under multiple discrimination.

The difference between Definition 4.6 and Definition 4.7 is subtle but central to detecting *prima facie* individual discrimination. In multiple discrimination, we require c to be discriminated *separately* q times as a member of each protected attribute it belongs to: e.g., as a female and as a non-white individual. In intersectional discrimination, instead, we require c to be discriminated *simultaneously* as a member of all the q protected attributes

it belongs to: e.g., as a female-non-white individual. As we discuss below, this distinction has clear modeling implications. The tension between these types of multidimensional discrimination occurs as it is possible for c not to suffer multiple discrimination while suffering intersectional discrimination (Crenshaw, 1989). This is, in particular, troubling as only the former is recognized under EU non-discrimination law (Xenidis, 2020).

For the present k-NN CST implementation we operationalize the two forms of multidimensional discrimination as follows:

- For multiple discrimination we run CST separately for each A_i, including the generation of the corresponding counterfactual datasets via each $do(A_i := 0)$; and look for individual cases in which discrimination is detected across all runs.
- For intersectional discrimination we create the *intersectional protected attribute* A^* as in Definition 4.7; generate the corresponding counterfactual dataset via $do(A^* := 0)$; and run a single CST as we would for $|A| = 1$.

Beyond using Definitions 4.6 and 4.7, respectively, both procedures have implications on Section 4.2. For multiple discrimination, we repeat Definitions 4.1, 4.2, and 4.3 for each of the q protected attributes. For intersectional discrimination, once we have generated A^*, we apply only once Definitions 4.1, 4.2, and 4.3 for this "new" protected attribute. Section 5.3 showcases both of these procedures for testing multidimensional discrimination. Additionally, under these two producers we can also explore positive discrimination for multiple and intersectional discrimination using k-NN CST. We would simple revisit, respectively, Definitions 4.6 and 4.7 by considering the opposite effect, meaning the relevant delta(s) being less than τ. We leave this for future work.

5. Experiments

We now showcase CST using its k-NN implementation (k-NN CST). Throughout this section, we compare it to its standard counterpart (k-NN ST) of Thanh et al. (2011) and counterfactual fairness (CF) of Kusner et al. (2017). Henceforth, we drop the "k-NN" for CST and ST. As noted in Remark 4.2, we exclude the search centers when comparing CST to ST (i.e., CST w/o) and include them (i.e., CST w/) when comparing CST to CF.[10]

5.1 Setup

We use a significance level of $\alpha = 0.05$, an accepted deviation of $\tau = 0.0$, and the neighborhood sizes of $k \in \{15, 30, 50, 100, 250\}$. In practice, we expect these parameters to be provided by the legal context. For instance, $k = 30$ is a common size used in non-algorithmic situation testing (Rorive, 2009). See Appendix C for additional experiments.

Section 5.2 focuses on comparing CST to ST and CF, while Section 5.3 focuses on illustrating CST for multidimensional discrimination testing. We use synthetic and real data, each describing high-stakes decision-making scenarios involving an ADM $b()$. We assume a corresponding SCM $\mathcal{M}$ and DAG $\mathcal{G}$ for each scenario. $\mathcal{M}$ and $\mathcal{G}$ are only needed for CST and CF. Similar to α, τ, and k, we expect $\mathcal{M}$ and $\mathcal{G}$ to be provided. The assumptions

10. The code is available in this repository: https://github.com/cc-jalvarez/counterfactual-situation-testing.

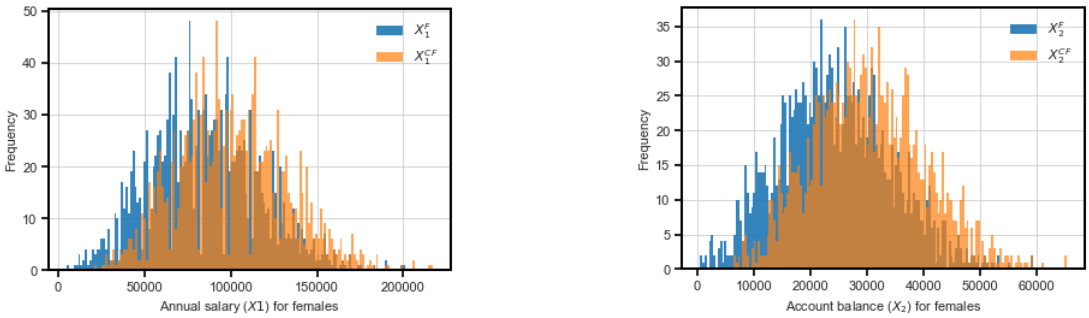


Figure 2: Distribution of the attributes used by the ADM $b()$ in Section 5.2. The histograms compare the $\mathcal{D}$ and $\mathcal{D}^{CF}$ datasets for the female applicants. The shift to the right of both X_1^{CF} and X_2^{CF} shows the negative impact of A on X_1 and X_2.

on $\mathcal{M}$ and $\mathcal{G}$ in Section 3 simplify the abduction step when generating $\mathcal{D}^{CF}$. We stress that these assumptions are convenient but not necessary.

5.2 An Illustrative Example

Let us consider the loan application scenario introduced in Example 1.1. We create the synthetic dataset $\mathcal{D}$ based on Figure 1. It is a modified version of Karimi et al. (2021, Figure 1). We add the protected attribute gender (A), which affects an applicant's annual salary (X_1) and bank balance (X_2) used by the bank's ADM $b()$ for approving ($\widehat{Y} = 1$) or rejecting ($\widehat{Y} = 0$) a 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. We define f_1 as $X_1 \leftarrow (-\$1500)\cdot\text{Poi}(10)\cdot A + U_1$ and f_2 as $X_2 \leftarrow (-\$300)\cdot\mathcal{X}^2(4)\cdot A + (3/10)\cdot X_1 + U_2$, assuming $U_1 \sim \$10000\cdot\text{Poi}(10)$ and $U_2 \sim \$2500\cdot\mathcal{N}(0,1)$. We define $b()$ as $\widehat{Y} = \mathbb{1}\{X_1 + 5\cdot X_2 > \$225000\}$. Importantly, with Figure 1 we have access to the data generating model of $\mathcal{D}$. It allows to study CST when we are able to control for potential model misspecifications in the SCM $\mathcal{M}$. $\mathcal{D}$ represents a biased scenario in which female applicants are penalized systematically in X_1 and X_2. 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 jobs (Criado-Perez, 2019). $\mathcal{D}$ represents a non-neutral status quo and, thus, a setting of interest to indirect non-discrimination law (Wachter et al., 2020).

We study the behavior of $b()$ toward A in $\mathcal{D}$. We use CST, ST, and CF to answer whether $b()$ discriminates against female loan applicants. For CST and CF, we generate $\mathcal{D}^{CF}$ using Figure 1 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. These effects can be seen in Figure 2 by the rightward shift experienced in X_1 and X_2 for all female applicants when going from the factual to the counterfactual world. $\mathcal{D}$ has a total of 1712 females (34.29%) and, overall, $b()$ has an acceptance rate of 53.5%. Using *demographic parity* (DP) (Kamiran & Calders, 2009), we also observe the unfairness of $b()$ with 13.5% acceptance probability for female and 40% for male applicants.[11]

11. Formally, we define DP as $P(\hat{Y}|A = 1) = P(\hat{Y}|A = 0)$. We only use DP as we do not have access to the ground-truth Y and cannot use, e.g., equalized odds (Hardt et al., 2016).

Table 1: Number and (% w.r.t. females) of individual discrimination cases based on A in Figure 1. Marked by * are the number of statistically significant cases.

Method	$k = 15$	$k = 30$	$k = 50$	$k = 100$	$k = 250$
CST w/o	288 (16.8%)	313 (18.3%)	342 (20%)	395 (23.1%)	534 (31.2%)
	272* (15.9%)	306* (17.9%)	331* (19.3%)	383* (22.4%)	519* (30.3%)
ST	55 (3.2%)	65 (3.8%)	84 (5%)	107 (6.3%)	204 (11.9%)
	44* (2.6%)	57* (3.3%)	65* (3.8%)	85* (5%)	148 (8.6%)
CST w/	420 (24.5%)	434 (25.4%)	453 (26.5%)	480 (28%)	557 (32.5%)
	272* (15.9%)	307* (17.9%)	334* (19.5%)	385* (22.5%)	520 (30.4%)
CF	376 (22%)	376 (22%)	376 (22%)	376 (22%)	376 (22%)
	241* (14.1%)	253* (14.8%)	265* (15.5%)	288* (16.8%)	352* (20.6%)

Table 1 reports the results for all methods based on Definitions 4.4 and 4.5. Under Definition 4.4, we detect individual discrimination when the complainant's $\Delta p > \tau$; and under Definition 4.5, we detect individual discrimination that is statistically significant given α when $\Delta p > \tau$ and τ does not fall within Δp's CI. Figures 4 and 5 further show how the results change under these definitions for all methods as k increases, including cases beyond $k = 250$. Let us discuss these results in detail.

5.2.1 CST Relative to ST

We consider CST w/o as ST excludes the search centers when testing for individual discrimination. What is clear from Table 1 is that CST w/o detects more cases than ST across all k sizes, even when accounting for statistical significance. The results highlight the impact of operationalizing *fairness given the difference* since the only difference between the two methods is the choice of the test search center. ST performs an idealized comparison. If we consider the tuple $(x_1 = 35000, x_2 = 7948, a = 1)$ as complainant c, then ST builds the test group by finding the k most similar male profiles under d (3) using c as the test search center. CST w/o, conversely, performs a more flexible comparison under *fairness given the difference*. For the same c and d (3), it instead uses the counterfactual tuple $(x_1^{CF} = 50796, x_2^{CF} = 13852, a^{CF} = 0)$ as the test search center. The control group is the same for both ST and CST as each uses c as the control search center.

Figure 3 shows the impact of *fairness given the difference*. We randomly choose five complainants that are discriminated by $b()$ according to ST and CST w/o for $k = 15$, and plot the distributions of X_1 and X_2 as boxplots for the control group (ctr), ST test group (tst-st), and CST w/o test group (tst-cf). The tst-cf are above the tst-st boxplots, indicating male profiles that are likelier to be accepted by $b()$. As all 55 ST cases for $k = 15$ are included in the 288 CST w/o cases, the difference in test groups explains the difference in the number of cases between the methods. Tables 2 and 3 illustrate this point.

Table 2 reports the average summary statistics for the cases detected by ST and CST w/o for $k = 15$. Notice that (the average of) the average and standard deviation for X_1 and X_2 are similar between the ST control and test groups, while such similarity does not

Table 2: Summary statistics for cases detected by ST and CST w/o for $k = 15$. We present the averages for the control groups (ctr's) and the corresponding test groups (tst's).

Average of	ctr's for ST and CST w/o	tst's for ST	tst's for CST w/o
Num. of neg. $\hat{Y}$	8.82	2.22	0.00
Prp. of neg. $\hat{Y}$	0.59	0.15	0.00
Avg. Annual salary	94372.12	96181.82	106569.70
Std. Annual salary	1646.29	611.57	286.34
Avg. Account balance	26092.93	26347.46	30141.28
Std. Account balance	558.30	352.78	273.45

Table 3: Summary statistics, similar to Table 2 but for all cases detected by CST w/o only. We include the corresponding ST test groups for comparison.

Average of	ctr's for ST and CST w/o	tst's for ST	tst's for CST w/o
Num. of neg. $\hat{Y}$	13.74	14.79	0.78
Prp. of neg. $\hat{Y}$	0.92	0.99	0.05
Avg. Annual salary	86332.47	85911.30	96898.43
Std. Annual salary	1524.33	906.11	391.49
Avg. Account balance	24734.11	24790.05	28677.11
Std. Account balance	482.15	152.07	171.61

occur between the CST w/o groups. The CST w/o test groups show a higher annual salary and account balance than their ST counterparts. It translates into a lower average number and proportion of negative $\hat{Y}$ as these male profiles are likelier to be accepted by $b()$. There is still a clear difference in outcomes between the control groups and the ST and CST w/o test groups, which leads to both methods detecting these cases.

Table 3 reports the same summary statistics but for cases detected by CST w/o only. For comparison, we include the corresponding ST test groups. The groups for ST are again similar (and lower in average values) between them in terms of X_1 and X_2, but are also similar in terms of the number of negative $\hat{Y}$, explaining why ST does not detect these cases. The CST w/o test groups, instead, exhibit almost no cases of negative $\hat{Y}$. Intuitively, ST is comparing females and males with similar applicants not suitable for a loan while CST w/o is comparing non-suitable female to suitable male applicants. This difference between ST and CST w/o is stark because $\mathcal{D}$ is purposely generated with a systematic bias against female applicants. We expect the CST w/o test groups to be more suitable than their ST counterparts as the former's test search centers account for the downstream effects of A on X_1 and X_2 while the latter's test search centers ignore any effect at all.

Figure 4 illustrates how CST w/o and ST differ in terms of number of discrimination cases and Δp up until $k = 500$ for all and statistically significant cases. In both plots, CST w/o and ST follow a similar trend and with signs of a slow convergence toward the other. The impact of the *mutatis mutandis* over the *ceteris paribus* manipulation occurs on smaller k's and persists over larger k's, clearly differentiating CST w/o from ST and

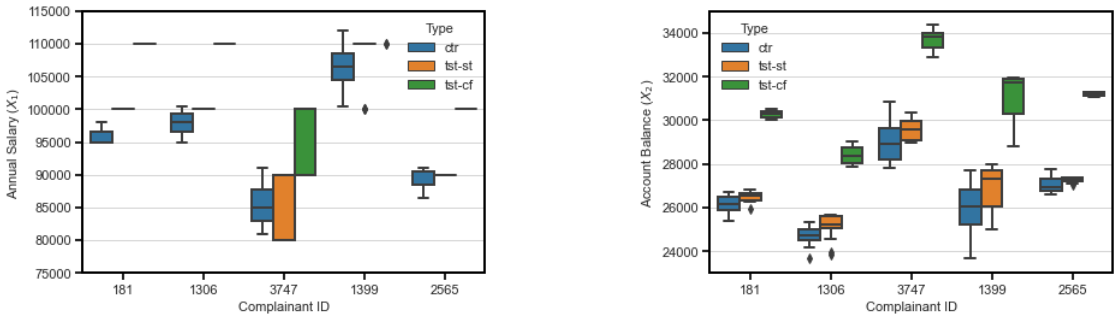


Figure 3: The boxplots for a subset of cases detected by ST and CST w/o for $k = 15$. We compare the ST and CST w/o control group (ctr) versus the ST (tst-st) and CST w/o (tst-cf) test groups. The tst-st are closer to ctr than tst-cf, illustrating the *fairness given the difference* behind CST and the *idealized comparison* behind ST.

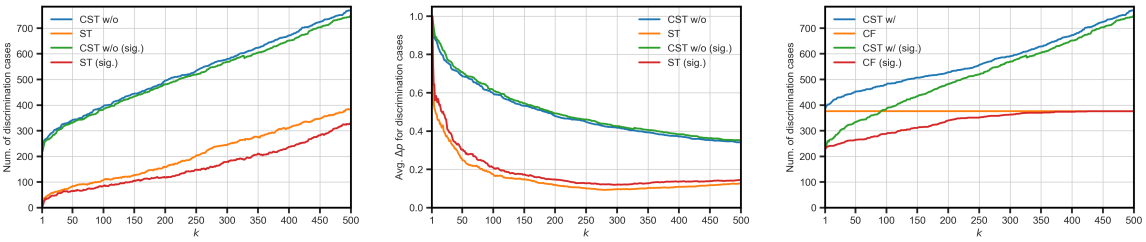


Figure 4: Left and center: Number of cases by CST w/o and ST and their avg. Δp. Right: Number of cases by CST w/ and CF. We plot all and statistically significant (sig.) cases.

its operationalization of *fairness given the difference* for sensible ranges of k. The higher number of cases by CST w/o over ST (left) is driven by test groups that are, on average, dissimilar to the control groups, leading to a larger average Δp by CST w/o over ST (center). Both methods show similar trends when considering all and statistically significant cases, meaning the difference between CST w/o and ST is statistically significant. The results validate the use of CST w/o over ST as well as the viability of the *mutatis mutandis* over the *ceteris paribus* manipulation for testing discrimination.

5.2.2 CST Relative to CF

We consider the CST w/ as CF uses the search centers for measuring the counterfactual fairness of $b()$. Back to the illustrative factual and counterfactual tuples of c, we now include the control search center $(x_1 = 35000, x_2 = 7948, a = 1)$ and the test search center $(x_1^{CF} = 50796, x_2^{CF} = 13852, a^{CF} = 0)$ into the k-NN neighborhoods to be compared for c. These are the two instances we compare under CF to test for the counterfactual fairness of $b()$. For comparison, we define CF discrimination as a case in which the factual $\widehat{y}_c = 0 \in \mathcal{D}$ (from negative outcome) becomes $\widehat{y}_c^{CF} = 1 \in \mathcal{D}^{CF}$ (to positive outcome) after the intervention on A. This definition aligns with using $\tau = 0.0$ for CST w/.

Table 1 shows how CST w/ detects for each k a higher number of cases than CF, even when accounting for statistical significance. In fact, all cases detected by CF are

Table 4: Subset of cases detected by both CST w/ and CF for $k = 15$. The * denotes statistical significance based on the one-sided CI.

Comp. (ID)	p_c	p_t	Δp	CI (10)	CI (11)
44	1.00	0.00	1.00*	[1.00, $+\infty$)	[1.00, 1.00]
55	0.81	0.00	0.81*	[0.65, $+\infty$)	[0.62, 1.00]
150	1.00	0.94	0.06	[-0.04, $+\infty$)	[-0.06, 0.18]
203	1.00	0.87	0.13	[-0.01, $+\infty$)	[-0.04, 0.29]
218	0.56	0.00	0.56*	[0.36, $+\infty$)	[0.32, 0.81]

Table 5: Subset of cases detected by CST w/ and not CF for $k = 15$. The * denotes statistical significance based on the one-sided CI.

Comp. (ID)	p_c	p_t	Δp	CI (10)	CI (11)
5	0.06	0.00	0.06	[-0.04, $+\infty$)	[-0.06, 0.18]
147	0.50	0.00	0.5*	[0.29, $+\infty$)	[0.26, 0.75]
435	0.38	0.00	0.38*	[0.18, $+\infty$)	[0.14, 0.61]
1958	0.13	0.00	0.13	[-0.01, $+\infty$)	[-0.04, 0.29]
2926	0.75	0.00	0.75*	[0.57, $+\infty$)	[0.54, 0.96]

also detected by CST w/. CF is independent from k as it applies only to the individual comparison of the factual and counterfactual tuples for a given c. However, the CI used for measuring the statistical significance of CF discrimination is a function k and, thus, varies with it (cfr., (10) and (11)). As k increases, so does the number of statistically significant CF and CST w/ individual discrimination cases. Both CST w/ and CF use the one-sided CI (10) as we test for $\Delta p > \tau$. Figure 4 (right) further shows how CST w/ and CF evolve as k increases up to $k = 500$ in terms of the number of cases detected. It aligns with Table 1. We do not plot the average Δp as CF discrimination focuses on the literal comparison of the factual and counterfactual instances of c. Further, the number of significant CF cases is bounded by all CF cases. As long as CST w/ detects more cases than CF, it appears that all CF cases become statistically significant under a large enough k.

What sets CST w/ apart from CF is twofold. First, CST w/ equips the CF comparison with certainty measures. Table 4 shows individual cases of discrimination detected by both CF and CST w/ for $k = 15$ along with the one-sided (10) and two-sided (11) CI. The latter CI is specific to CF and it is built through the CST w/ k-NN implementation. Beyond the CF discrimination claim, this CI provides a measure of certainty to the factual versus counterfactual comparison behind CF. Hence, CST complements CF beyond detecting discrimination. Second, CST w/ detects cases of individual discrimination that are counterfactually fair. In Table 5 reports cases for $k = 15$ that do not exhibit CF discrimination but exhibit a discriminatory pattern when looking at Δp. The results highlight the importance of requiring multiple comparisons to insure that the complainant's experience is representative of the discrimination claim.

Table 6: Summary statistics for cases detected by both CF and CST w/ under $k = 15$. We present averages for the control groups (ctr's) and the test groups (tst's).

Average of	ctr's CST w/	tst's for CST w/
Num. of neg. $\hat{Y}$	15.46	5.75
Prp. of neg. $\hat{Y}$	1.03	0.38
Avg. Annual salary	83281.25	94323.46
Std. Annual salary	1410.97	1564.37
Avg. Account balance	23752.64	27762.60
Std. Account balance	449.85	550.18

Table 7: Summary statistics for cases detected by CST w/ under $k = 15$ and not CF. We present averages for the control groups (ctr's) and the test groups (tst's).

Average of	ctr's CST w/	tst's for CST w/
Num. of neg. $\hat{Y}$	5.23	0.00
Prp. of neg. $\hat{Y}$	0.35	0.00
Avg. Annual salary	92892.05	104541.25
Std. Annual salary	1592.33	1434.18
Avg. Account balance	26890.63	31161.38
Std. Account balance	509.61	545.73

Tables 6 and 7 report how a counterfactually fair $b()$ still discriminates according to CST w/. As argued in Section 4.4, it occurs when considering a complainant and its counterfactual that are close to and on the same side of the decision boundary of $b()$. When building the neighborhoods, CST w/ may include profiles from both sides of the decision boundary. Table 6 summarizes the group characteristics of the cases detected by CF and CST w/. If we consider (the average of) the average X_1 and X_2 for control and test groups, we obtain that $b()$ rejects the former and accept the latter, meaning it is not counterfactually fair. Table 7, instead, summarizes cases detected by CST w/. For those cases, using again (the average of) the average X_1 and X_2 for control and test groups, $b()$ accepts both.

5.2.3 CST w/o and CST w/

Finally, we consider the two CST versions. Three patterns are clear in Table 1. First, CST w/ detects a higher number of individual discrimination cases than CST w/o for all k. Second, this difference decreases between the CST versions as k increases. Third, CST w/ and CST w/o detect roughly the same number of cases when accounting for statistical significance. Let us explore these patterns. Note that $\tau = 0.0$; any deviation of p_c from p_t (read, Δp) is detected by CST. Such deviation is statistically significant (read, representative) depending on the composition of the control and test groups behind p_c and p_t.

What differentiates CST w/o from w/ is the exclusion of the search centers (Remark 4.2), making the size of the neighborhoods under CST w/o of size k and under CST w/ of size $k + 1$. Back to the illustrative factual and counterfactual tuples of c, the CST w/o control

Table 8: Summary statistics for CST w/o cases for $k = 15$ for the control (ctr's) and test groups (tst's) split by statical significance, which we denote with *.

Average of	ctr's	tst's	ctr's*	tst's*
Num. of neg. $\hat{Y}$	3.94	2.62	13.32	0.51
Prp. of neg. $\hat{Y}$	0.26	0.17	0.89	0.03
Avg. Annual salary	92808.34	104750.00	87577.21	98392.16
Std. Annual salary	1919.18	1182.73	1525.77	323.69
Avg. Account balance	25923.28	30121.55	24938.92	28888.21
Std. Account balance	652.67	290.97	487.52	185.18

Table 9: Summary statistics for CST w/ cases for $k = 15$ for the control (ctr's) and test groups (tst's) split by statical significance, which we denote with *.

Average of	ctr's	tst's	ctr's*	tst's*
Num. of neg. $\hat{Y}$	14.72	13.66	14.21	0.51
Prp. of neg. $\hat{Y}$	0.92	0.86	0.89	0.03
Avg. Annual salary	78240.29	89308.43	87578.81	98705.10
Std. Annual salary	1331.29	1672.54	1483.66	1484.45
Avg. Account balance	22503.79	26478.22	24939.78	29011.26
Std. Account balance	422.47	554.67	474.41	547.02

and test group are the same as the CST w/ ones with the key difference that the CST w/ control group has a *(k+1)-th* that is $(x_1 = 35000, x_2 = 7948, a = 1)$ and the test group has a *(k+1)-th* that is $(x_1^{CF} = 50796, x_2^{CF} = 13852, a^{CF} = 0)$. The additional tuple pair compared in CST w/ is, notably, the most important pair as the complainant and its counterfactual denote the closest possible comparison under CST. Clearly, the inclusion (or exclusion) of the complainant-counterfactual pair into Δp of CST w/ (or CST w/o) drives the statistically insignificant difference in the number of detected cases. Based on Table 1, such pair is able to deviate Δp from τ but is not representative of the corresponding control and test neighborhoods.

Let us consider the two CST versions for $k = 15$. Tables 8 and 9 summarize the control and test groups, respectively, of CST w/o and CST w/ split by statistical significance. These tables are not mutually exclusive as CST w/o and CST w/ detect the same 272 statistically significant cases and the 288 CST w/o cases are included in the 420 CST w/ cases in Table 1. The group pairs within each table are mutually exclusive as, different from previous tables, we split between statistically significant and non-significant cases. Based on the average characteristics in both tables, these groups are almost identical and enjoy, on average, a large Δp's. Together, it explains why the inclusion (exclusion) of the the complainant and its counterfactual from, respectively, the control and test group in CST w/ (CST w/o) has little impact on these cases. Comparing now the significant versus non-significant control and test groups in each table, we observe that the latter have a smaller difference between

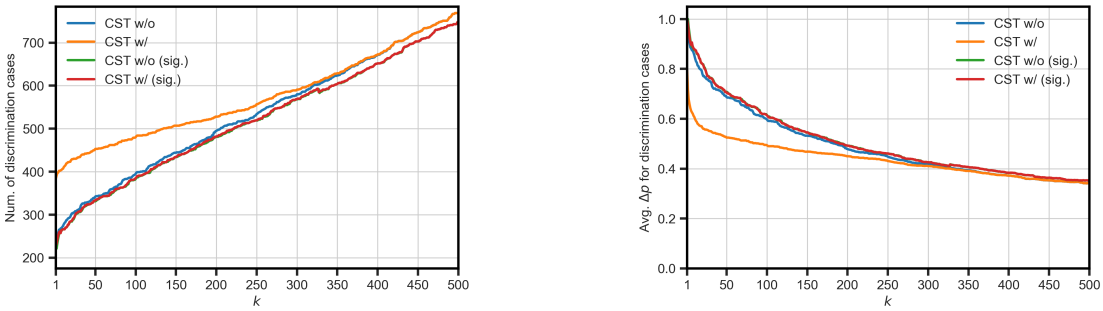


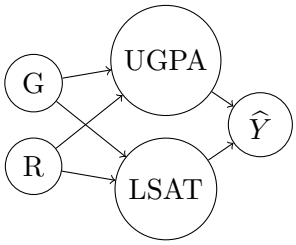
Figure 5: Number of cases by CST w/o and w/ and their average Δp, respectively. We plot all and statistically significant (sig.) cases for both versions of CST.

the average number of negative $\hat{Y}$ per group, which explains why these cases are detected by CST w/o and w/ but do not amount to statistically significant cases.

The non-statistically significant cases in Tables 8 and 9 vary more than the statistically significant cases in terms of their composition, with CST w/ having a considerably larger average number of negative decisions. This difference illustrates the role of including/excluding the search centers and their representativeness relative to their corresponding neighborhoods. Here, recall that CST w/ almost doubles CST w/o for all cases. CST w/ detects cases with more heterogeneous neighborhoods in terms of applicant suitability, which would explain the lower average proportion of negative outcomes. CST w/, instead, detects more homogeneous neighborhoods marked by a lack of suitability due to the large average proportion of negative outcomes. What occurs for the non-statistically significant CST w/ groups is that the complainant and its counterfactual disagree, meaning the former is rejected and the latter is accepted, which would be enough to deviate Δp from τ while both control and test neighborhoods appear to be, on average, clearly non-suitable.

What about for larger neighborhood sizes? In Table 1, CST w/o and CST w/ converge in number of cases as k increases. We argue that this occurs as larger neighborhoods diminish the impact of the *(k+ 1)-th* pair, meaning the CST w/ is, in fact, converging toward CST w/o. In Figure 5 we plot the effect of increasing k up to 500 on the number of discrimination cases and their average Δp. We do so for both CST versions, differentiating between all and statistically significant cases. These plots align with Table 1. The plots support the previous analysis for Tables 8 and 9 regarding the impact of the additional pair in CST w/. Consider the plots for the number of discrimination cases (left). As k increases, the number of cases increases for all four. Notably, CST w/o's all and statistically significant cases and CST w/'s statistically significant cases align throughout k, while CST w/'s all cases converges to these three. The same pattern occurs when we look at the plots for average Δp (right), though the average Δp decreases as k increases.

Given Table 1 and Figure 5, it is worth asking whether having two versions of CST is useful? Each version targets a different method, CST w/o relative to ST and CST w/ relative to CF; however, we might prefer CST w/o if expected to provide statistically significant results. What matters is that the difference between ST and both CST versions holds, highlighting the impact of using a *mutatis mutandis* over a *ceteris paribus* manipulation when testing for discrimination. Also important is that we are able to provide certainty



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

$$\widehat{Y} = b(UGPA, LSAT)$$

Figure 6: The auxiliary causal knowledge for Section 5.3. Let R denote race, G gender, $LSAT$ law school admissions test scores, $UGPA$ undergraduate grade-point average, and $\hat{Y}$ the admissions decision by $b()$.

measures around CF under both CST versions as long as k is large enough for Δp under CST w/ to converge to Δp under CST w/o.

Finally, the discrepancy between the CST versions raises questions on what it means for a Δp to be representative. Consider that we have not constrained $\mathcal{D}^{CF}$ w.r.t. $\mathcal{D}$, meaning it is possible for the counterfactual instances to deviate considerably from the factual instances. In that case, we would have "unrepresentative" counterfactual instances that motivate statistically insignificant comparisons, but that are still "unavoidable" in the sense that these same instances embody scenarios that should have happened for the complainants according to a SCM $\mathcal{M}$. We believe that judging the representativeness of the counterfactual world by using the factual world as the underlying population might not properly capture the substantive equality goals behind non-discrimination law (Wachter et al., 2020). We come back to this last discussion in Section 6.

5.3 Law School Admissions

Let us now consider the law school admissions scenario popularized by Kusner et al. (2017, Figure 2). We use US data from the Law School Admission Council survey (Wightman, 1998), and recreate an admissions scenario for a top US law school. We consider as protected attributes an applicant's gender (G, male/female), and race (R, white/non-white). We add the ADM $b(UGPA, LSAT) = \hat{Y}$, which considers the applicant's undergraduate grade-point average ($UGPA$) and law school admissions test scores (LSAT). If an applicant is successful, $\widehat{Y} = 1$; otherwise $\widehat{Y} = 0$. We summarize the scenario in Figure 6. We define the ADM $b()$ using the median entry requirements for the top US law school to derive the cutoff ψ.[12] Formally, we define $b()$ as $\mathbb{1}\{(0.6 \cdot UGPA + 0.4 \cdot LSAT) > \psi\}$. The cutoff is the weighted sum of 60% in $UGPA$ (3.93 over 4.00), and 40% $LSAT$ (46.1 over 48), giving a total of 20.8; the maximum possible score given $b()$ is 22. The SCM $\mathcal{M}$ and DAG $\mathcal{G}$ follow Kusner et al. (2017), with b_U and b_L denoting the intercepts; β_1, β_2, λ_1, λ_2 the weights; and $U_1 \sim \mathcal{N}$ and $U_2 \sim$ Poi the probability distributions.

We study the behavior of $b()$ toward G and R. The dataset $\mathcal{D}$ contains $n = 21790$ applicants, 43.8% being female, 16.1% being non-white, and 8.4% being non-white-female.

12. That being Yale University Law School; see https://www.ilrg.com/rankings/law/index/1/asc/Accept

Table 10: Number (and % w.r.t. non-whites) of individual discrimination cases based on R using Figure 6. Marked by * are the statistically significant cases.

Method	$k = 15$	$k = 30$	$k = 50$	$k = 100$	$k = 250$
CST w/o	256 (7.3%) 244* (6.9%)	309 (8.8%) 301* (8.6%)	337 (9.6%) 323* (9.2%)	400 (11.4%) 391* (11.2%)	503 (14.4%) 494* (14.1%)
ST	33 (0.9 %) 28* (0.8%)	51 (1.5%) 28* (0.8%)	61 (1.7%) 45* (1.3%)	64 (1.8%) 47* (1.3%)	78 (2.2%) 61 (1.7%)
CST w/	286 (8.2%) 244* (6.9%)	309 (8.8%) 301* (8.6%)	337 (9.6%) 323* (9.2%)	400 (11.4%) 391* (11.2%)	503 (14.4%) 494* (14.1%)
CF	231 (6.6%) 190* (5.4%)	231 (6.6%) 231* (6.6%)	231 (6.6%) 231* (6.6%)	231 (6.6%) 231* (6.6%)	231 (6.6%) 231* (6.6%)

Table 11: Number (and % w.r.t. females) of individual discrimination cases based on G Figure 6. Marked by * are the statistically significant cases.

Method	$k = 15$	$k = 30$	$k = 50$	$k = 100$	$k = 250$
CST w/o	78 (0.8%) 43* (0.5%)	120 (1.3%) 88* (0.9%)	253 (2.7%) 160* (1.7%)	296 (3.1%) 221* (2.3%)	493 (5.2%) 341* (3.6%)
ST	77 (0.8%) 57* (0.6%)	101 (1.1%) 69* (0.7%)	229 (2.4%) 111* (1.2%)	258 (2.7%) 124* (1.3%)	484 (5.1%) 366 (3.8%)
CST w/	99 (1.0%) 54* (0.6%)	129 (1.4%) 92* (0.9%)	267 (2.8%) 160* (1.7%)	296 (3.1%) 221* (2.3%)	493 (5.2%) 341* (3.6%)
CF	56 (0.6%) 20* (0.2%)	56 (0.6%) 15* (0.2%)	56 (0.6%) 30* (0.3%)	56 (0.6%) 21* (0.2%)	56 (0.6%) 32* (0.3%)

Despite $b()$ being externally imposed by us for the purpose of illustrating the CST framework, under $b()$ only 0.8% of the female applicants are successful compared to 1.5% of the male applicants; similarly, only 0.2% of the non-white applicants are successful compared to 2.2% of the white applicants. It is also the case when considering the intersectional group of non-white-females, with only 0.06% of these applicants being admitted to law school based on $b()$ compared to the 2.25% of white-female, non-white-male, and white-male successful applicants. Notice that $b()$ is highly selective, with an acceptance rate of just 2.31%, or 505 out of 21790 applicants considered. Using DP as a fairness metric, $b()$ would still be considered unfair toward female, non-white, and non-white-female applicants.

5.3.1 Single Discrimination

Does $b()$ discriminate against non-white applicants? To answer this question using CST and CF, we generate the corresponding $\mathcal{D}_R^{CF}$ using Figure 6 based in the intervention $do(R := 0)$, or *what would have happened had all law school applicants been white?* Similarly, does $b()$

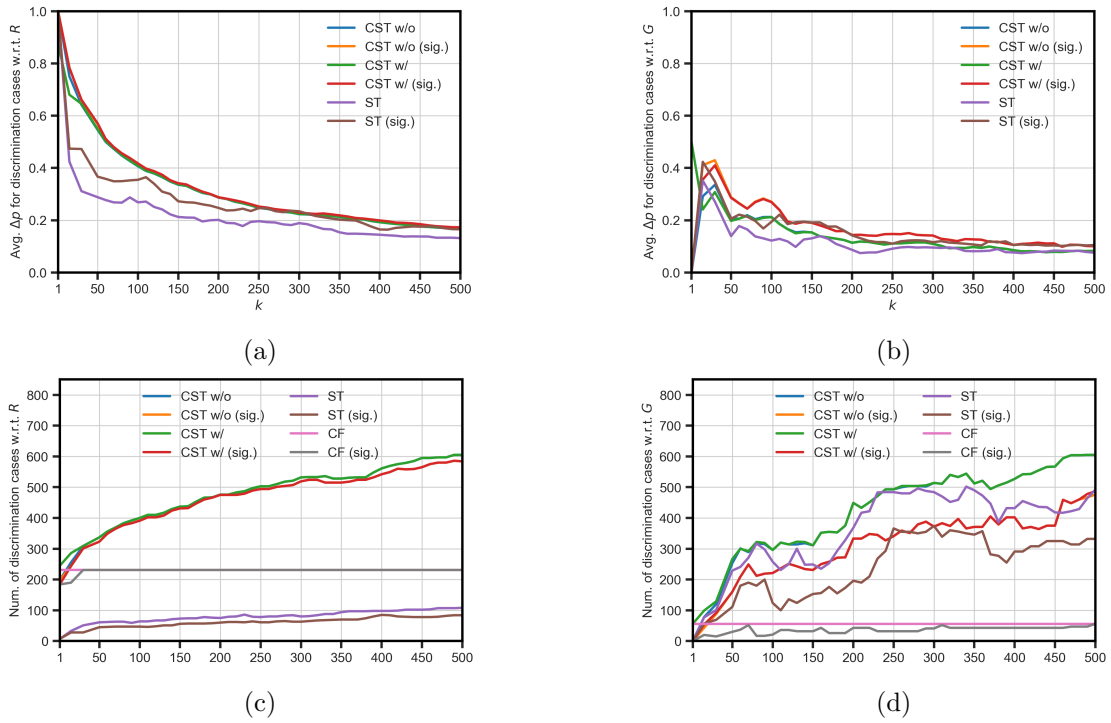


Figure 7: Average Δp and number of cases for race (R) and gender (G), respectively. We plot all and statistically significant (sig.) cases for each method.

discriminate against female applicants? To answer this question using CST and CF, we generate the corresponding $\mathcal{D}_G^{CF}$ using Figure 6 based in the intervention $do(G := 0)$, or *what would have happened had all law school applicants been male?* Both questions share the same $\mathcal{D}$. Similar to Section 5.2, we use Definition 4.4 for detecting individual discrimination cases and Definition 4.5 for determining whether these cases are statistically significant. Tables 10 and 11 show the results for all methods.

Tables 10 and 11 report similar patterns to those in Table 1. CST w/o detects a higher number of cases than ST; CST w/ detects a high number of cases than CF; and CST w/ detects a higher number than CST w/o. For all these patterns, though, the differences between the corresponding methods is smaller than those observed in Table 1. This is due to the composition of $\mathcal{D}$ and the nature of $b()$. Here, we work with a larger $\mathcal{D}$ (21790 versus 5000 applicants) and a more selective $b()$ (2.31% versus 53.3% acceptance rate). CST w/o and CST w/ in Table 11 converge already at larger neighborhood sizes, which does not occur in Table 1 (there is, though, a clear pattern that it occurs eventually as shown in Figure 5). We also observe similar patterns once we account for statistical significance with CST w/o and CST w/ reaching the same number of cases in both tables for $k = 250$.

Figure 7 supports Tables 10 and 11, showing the average Δp and the number of cases up to $k = 500$ for all methods.[13] For the average Δp, as shown in sub-figures (a) and (b), it decreases as k increases, with all methods seemingly converging to a single value. By

13. Due the size of $\mathcal{D}$, we run the methods for k equals 1, 15, 30 and between 50-500 in increments of 10.

observing that, as $k \to \infty$, the neighborhoods of the complainant and its counterfactual include all protected and unprotected instances, respectively, in $\mathcal{D}$, this value turns out to be the difference in DP: $P(\hat{Y}|A=1) - P(\hat{Y}|A=0)$ (cfr., footnote 11) . The average Δp for significant cases is higher to, see ST in (a), or almost the same as, see CST w/ in (a), for all the cases. As k increases, the control and tests groups constructed by ST, CST w/o, and CST w/ start considering new instances further away from the search centers and whatever detected initial deviation from τ dissipates slowly. Similarly, as shown in sub-figures (c) and (d), the number of cases increases as k increases except for CF which is independent of k and its significant cases that are bounded by CF itself. We observe the CST versions converging, especially when accounting for statistical significance, while ST remains below both CST versions in (c) and mimics CST w/o in (d). The number of significant cases is lower to ST in (c), or almost the same as both CST versions in (c) for all cases.

The number of cases varies across the methods between Tables 10 and 11. This is due to race and gender having different non-protected and protected search spaces. The results are comparable, but represent separate tests for single discrimination. Recall that non-whites represent 16.1% while females represent 43.8% of $\mathcal{D}$. It means that CST has access to a smaller search space when building the control groups for non-white complainants relative to female complainants. Notably, in Table 10 statistically significant CF cases reach the 231 CF total cases within the k values considered. This does not occur in Table 11 in which the statistically significant CF cases slowly increase, though in a non-monotonically way, toward the 56 CF total cases. Such oscillation, we believe, is due to the statistical estimator not yet reaching its asymptotic behavior. These are, though, minor fluctuations as the number of statistically significant cases is around 0.2-0.3%.[14] In Figure 7 we observe this non-monotonic increase for number of cases and decrease for the average Δp of cases detected more clearly in the sub-figures (a) and (c) for race than for the sub-figures (b) and (d) for gender. The plots for gender are considerably less smooth than those for race. The composition of $\mathcal{D}$ clearly plays a role here as females represent 43.8% and non-whites 16.1% of the dataset, with the k-NN based methods varying more between iterations as they explore a much denser search space.

5.3.2 Multidimensional Discrimination

We present the results for the forms of multidimensional discrimination, multiple and intersectional. Given the focus on gender and race, the non-protected group amounts to the non-protected groups based on race and gender: i.e., white and male applicants. As these two groups of applicants are not mutually exclusive, white-females and non-white-males are also part of the non-protected group. This point is clearer when we consider the intersection of race and gender and focus on the protected group that is non-white-female applicants: the complementary of such group, meaning the non-protected group, includes white-female, non-white-male, and white-male applicants.

Following Definition 4.6, we count as multiple discrimination based on race and gender when $\Delta p > \tau$ occurs separately for each of these protected attributes. Only those cases that are statistically significant for each protected attribute under CI (10)—given the Bonferroni

14. We use the CI (10) from CST w/ though conditioned on CF discrimination occurring. We only detect 56 cases of CF discrimination (Table 11). As k increases, we always look at these complainants.

Table 12: Number (and % w.r.t. non-white-females) of multiple individual discrimination cases in Section 5.3 for R and G. Marked by * are the statistically significant cases.

Method	$k = 15$	$k = 30$	$k = 50$	$k = 100$	$k = 250$
CST w/o	8 (0.44%) 4* (0.22%)	10 (0.55%) 6* (0.33%)	20 (1.09%) 11* (0.60%)	20 (1.09%) 17* (0.93%)	40 (2.18%) 24* (1.31%)
ST	5 (0.27%) 0* (0.0%)	5 (0.27%) 0* (0.0%)	12 (0.65%) 5* (0.27%)	19 (1.04%) 5* (0.27%)	24 (5.1%) 15* (0.82%)
CST w/	9 (0.49%) 4* (0.22%)	10 (0.55%) 9* (0.49%)	21 (1.15%) 11* (0.60%)	20 (1.09%) 17* (0.93%)	40 (2.18%) 24* (1.31%)
CF	5 (0.27%) 0* (0.0%)	5 (0.27%) 3* (0.16%)	5 (0.27%) 1* (0.05%)	5 (0.27%) 1* (0.05%)	5 (0.27%) 2* (0.11%)

Table 13: Number (and % w.r.t. non-white-females) of intersectional individual discrimination cases in Section 5.3 for $R \times G$. Marked by * are the statistically significant cases.

Method	$k = 15$	$k = 30$	$k = 50$	$k = 100$	$k = 250$
CST w/o	130 (7.1%) 130* (7.1%)	138 (7.5%) 138* (7.5%)	148 (8.1%) 148* (8.1%)	160 (8.7%) 160* (8.7%)	199 (10.9%) 199* (10.9%)
ST	14 (0.8%) 14* (0.8%)	14 (0.8%) 14* (0.8%)	17 (0.9%) 13* (0.7%)	24 (1.3%) 23* (1.3%)	29 (1.6%) 26* (1.4%)
CST w/	130 (7.1%) 130* (7.1%)	138 (7.5%) 138* (7.5%)	148 (8.1%) 148* (8.1%)	160 (8.7%) 160* (8.7%)	199 (10.9%) 199* (10.9%)
CF	113 (6.2%) 113* (6.2%)	113 (6.2%) 113* (6.2%)	113 (6.2%) 113* (6.2%)	113 (6.2%) 113* (6.2%)	113 (6.2%) 113* (6.2%)

corrected $\alpha/2$—amount to statistically significant cases. We still rely on the generated $\mathcal{D}_R^{CF}$ and $\mathcal{D}_G^{CF}$ for the construction of the test groups and the original $\mathcal{D}$ for the construction of the control groups. Does $b()$ discriminate against the none-white-female applicants as non-white *and* as female applicants? We present the results in Table 12. Figure 8, as shown in sub-figures (a) and (c), further illustrates the results up to $k = 500$. For the average Δp, we average those from cases detected separately under race and gender.

Following Definition 4.7, instead, we count intersectional discrimination based on race and gender when $\Delta p > \tau$ occurs for the intersection of these protected attributes. In practice, it means constructing the new protected attribute $R \times G$; updating $\mathcal{D}$ into $\mathcal{D}'$, such that $R \times G \in \mathcal{D}'$; and generating $\mathcal{D}'^{CF}$ given Figure 6 under $do(R \times G := 0)$. It implies a single discrimination run but under the "new" single attribute $R \times G$, representing the intersection of R and G. Cases are statistically significant under CI (10) based on $R \times G$. Formally, in Figure 6 we merge the R and G nodes into the single $R \times G$ in $\mathcal{G}$ node and do the same for the corresponding equations in $\mathcal{M}$ by interacting the dummy variables for R and G and re-estimating the regression weights (Wooldridge, 2015). Does $b()$ discriminate

against the none-white-female applicants? We present the results in Table 13. Figure 8, as shown in sub-figures (b) and (d), further illustrates the results up to $k = 500$.

In Tables 12 and 13, the three methods show similar patterns between them. CST w/o detects more cases relative to ST (including statistically significant cases); CST w/ detects more cases than CF (including statistically significant cases); and the two CST versions converge once statistical significance is considered. The same line of reasoning used before still applies here for understanding how CST, ST, and CF relate to each other. The difference is that the protected group and, in turn, the non-protected group are defined by more than one protected attribute. What is interesting in Table 12 is that CST w/o and CST w/ converge early on for all cases, not just for those cases that are statistically significant. We suspect these are cases that are clearly discriminatory under both race and gender, making them likely to be detected by multiple and intersectional discrimination testing. All cases in Table 13 are also statistically significant for all methods. It is due to $R \times G = 1$ representing the most un-favored protected group from the combination of R and G, which results in larger Δp's and a complete convergence of all and statistically significant cases for all methods relative to multiple discrimination. Compare, e.g., (a) versus (b) and (c) versus (d) in Figure 8. We discuss the last point in the next section.

5.3.3 On Multiple and Intersectional Discrimination

The results from the previous section support claims by legal scholars on the risk of not recognizing intersectional discrimination under non-discrimination law. These claims, to the best of our knowledge, date back to Crenshaw (1989) and have become prominent again with the ongoing discussion around algorithmic discrimination (Xenidis, 2020). We suspect that, since multiple and intersectional discrimination share the protected group and the non-protected groups, there is a tendency to dismiss the latter as a special case of the former. An in-depth legal discussion on the tension between multiple and intersectional discrimination is beyond this paper, but Tables 12 and 13 support the calls by these researchers to treat intersectional discrimination separate from multiple discrimination.

We acknowledge that, from a modeling perspective, a difference between Tables 12 and 13 is expected since we implement different procedures. Under multiple discrimination we look at the intersection of two separate single discrimination testing runs, while under intersectional discrimination we look at a single discrimination run representing the intersection. The claims by legal scholars like Crenshaw (1989) and Xenidis (2020), however, were not as apparent to us, meaning, in principle, we had no reason to expect a higher number of cases for intersectional discrimination over multiple discrimination. In fact, from a probability theory perspective (read, the conjunction rule), if we had to choose a difference between the Tables 12 and 13, we would have guessed the opposite, with multiple discrimination acting as an upper limit to intersectional discrimination. Given these results, we now would add that such expectation holds from a modeling perspective if we agree that the intersection of G and R is not its own category. Let us discuss further.

The two modeling procedures allow to represent the case in which $R \times G$ is its own category (intersectional) and in which is just the conjunction of R and G (multiple). This distinction, with legal origins as already argued (Crenshaw, 1989), in fact materializes through the counterfactual representation of each complainant under these two discrimina-

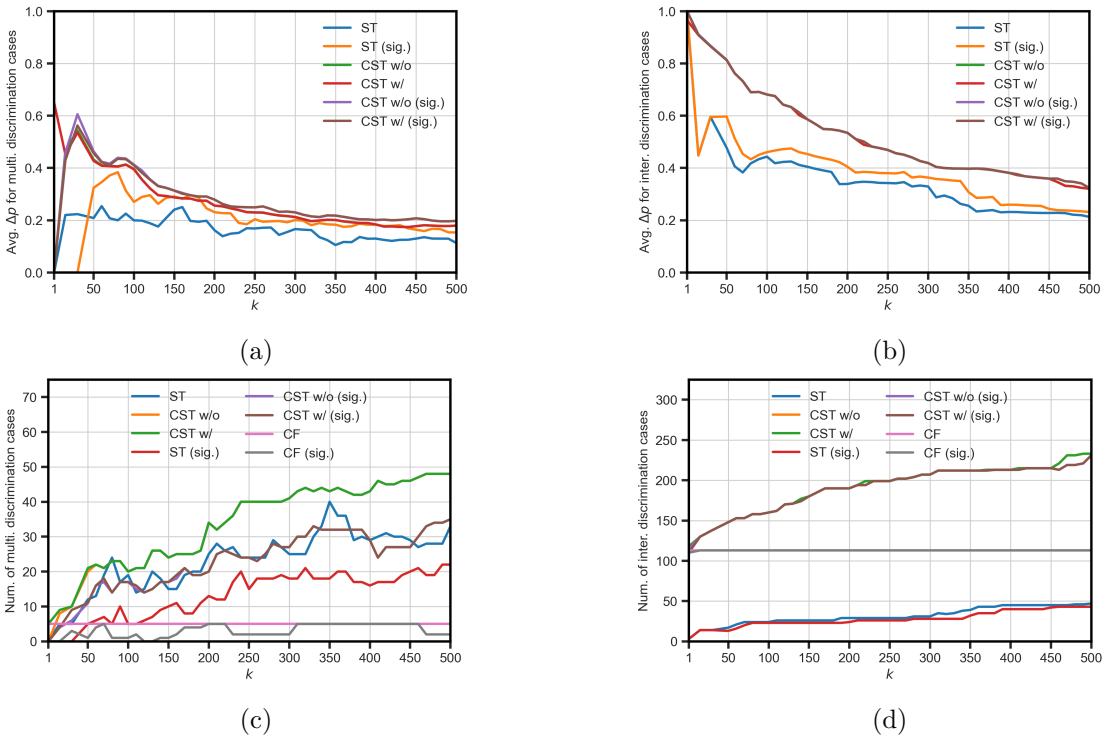


Figure 8: Avg. Δp and number of cases for multiple (left) and intersectional (right) discrimination on R and G. We plot all and statistically significant (sig.) cases for all methods.

tion testing procedures. Under multiple discrimination, we still rely on $\mathcal{D}_R^{CF}$ and $\mathcal{D}_G^{CF}$ for running the methods separately on R and G. We look at male and non-white conterfactuals separately. Although together they cover all the non-protected groups, they do not do so simultaneously: the complainant under this procedure would have been male or non-white, but not female-non-white. Under intersectional discrimination, instead, we rely on updated factual and counterfactual datasets based on a "new" protected attribute $R \times G$. The counterfactual for a given compliant implies simultaneously the possibilities of white-male, white-female, and non-white-male, which introduces more randomness.

In Figure 6, arguably, the worst-off sub-group between R and G is the group of non-white-female applicants. The logic here is that non-white males, meaning $R = 1$ and $G = 0$, can always resort to their gender and white females, meaning $R = 0$ and $G = 1$, can always resort to their race. Instead, non-white females have no single group within the space of $R \times G$ to resort to. When we test for multiple discrimination we allow for these movements to occur by looking at R and G separately. This is because, by not intersecting R and G, we do not consider the fact that the group at the intersection never has the choice to resort to a non-protected group. This lack of choice is what we represent when we test for intersectional discrimination by looking at $R \times G$ only. The modeling problem for these forms of multidimensional discrimination requires further research with the goal of formalizing the role of the intersection and how it influences the control and test search spaces while taking into account the legal considerations discussed.

Table 14: Avg. p_c and p_t for the ctr's and tst's groups of CST w/o for $k = 15$.

	Average					
	G's p_c	G's p_t	R's p_c	R's p_t	$R \times G$'s p_c	$R \times G$'s p_t
Multi. and inter.	0.59	0.33	0.51	0.00	0.66	0.00
Inter. only	0.96	0.94	0.93	0.23	0.93	0.01

Back to Tables 12 and 13, we argue that the observed difference comes from one protected attribute having a stronger influence than the other on the non-protected attributes. If that is the case, then testing separately for R and G should show individuals that are discriminated only by one of the protected attributes, which dismisses the multiple discrimination claim. Table 14 supports this argument. Notably, it is for this reason that lawyers discourage multiple discrimination claims and suggest that the complainant focuses on the most dominant protected attribute (Xenidis, 2020).

In Table 14, given the results from the previous section, we focus on CST w/o for $k = 15$ and look at individual discrimination cases detected as both multiple and intersectional discrimination. We report the average p_c and p_t for R, G, and $R \times G$. All multiple cases are included in the intersectional cases detected by CST w/o. The first row shows these multiple discrimination cases. We observe that the average p_c is greater than the average p_t, and thus the average $\Delta p > \tau$, for R, G, and $R \times G$. These are individual cases that suffer the negative effects of R and G, separately and simultaneously, when applying to law school under $b()$. The second row, instead, shows the intersectional cases only. For comparison, we provide the average p_c and p_t for these individuals' single discrimination tests for R and G. We observe that, on average, R is the dominant protected attribute with a considerable difference in the proportion of negative outcomes between the control and test groups. This is not the case for G where the average difference is negligible. Indeed, for these individuals, by looking at each protected attribute separately, we lose the multiple discrimination case. In doing so, we also lose focus on what occurs at the intersection of $R \times G$. The results in Table 14 capture this lack of movement between protected and non-protected statuses experienced by those individuals at the bottom of the intersection of R and G.

6. Discussion

With CST we move away from the standard, idealized comparison used in ST and other discrimination testing tools by operationalizing fairness given the difference. The results are promising but there are limitations to CST worth considering for future work. In this section, we discuss the main ones.

The reliability of $\mathcal{D}^{CF}$. In this work, we do not constrain the generation of the counterfactual dataset. In principle, it is not an issue as the counterfactual distribution that results from intervening a given SCM $\mathcal{M}$ is unique and is the closest possible parallel world to the factual world (Karimi et al., 2020; Woodward, 2005). The resulting $\mathcal{D}^{CF}$ is reliable conditional on what is assumed by the SCM $\mathcal{M}$. Therefore, if the auxiliary causal knowledge is biased, then so will the counterfactual distribution and, in turn, $\mathcal{D}^{CF}$. What amounts to

an unbiased SCM $\mathcal{M}$ is debatable when modeling humans behavior. We come back to this point in the next theme.

Even if we do have access to an unbiased SCM $\mathcal{M}$, however, it is still difficult to judge what amounts to a reliable $\mathcal{D}^{CF}$ when our point of reference is the factual $\mathcal{D}$. The issue here is accepting what $\mathcal{D}^{CF}$ represents based on our believes about $\mathcal{D}$. In Section 5.2, e.g., we find ourselves in a setting in which we have access to the data generating model of $\mathcal{D}$. The $\mathcal{D}^{CF}$, thus, represents the what would have been of the female applicants. In Table 1 we observe the impact of considering the complainant and its counterfactual, observing how using CST increases the number of cases detected relative to ST and how such cases are penalized once we consider the statistical significance of Δp. In Table 1, one could argue that these counterfactuals are not reliable as they lead to non-statistically significant cases. One could also, though, argue that it is counterintuitive and even counterproductive to measure the reliability of the counterfactual world using the factual world. In particular, when it comes to indirect discrimination testing and its goal of achieving substantive equality (Wachter et al., 2020), we argue that it is conceivable to generate counterfactual distributions that cannot be other than non-representative of the current, non-neutral status quo behind the factual world. This is the case in Section 5.2. For this reason alone, we expect *ceteris paribus* over *mutatis mutandis* to remain the preferred manipulation in discrimination testing as the idealized comparison is easier to motivate. Further research is needed for defining what we mean, or want to mean, when speaking of a reliable $\mathcal{D}^{CF}$ when testing for discrimination, especially, if we wish to implement the *mutatis mutandis* manipulation.

Auxiliary causal knowledge as groundtruth. Similarly, in this work we take as a given the derivation of the SCM $\mathcal{M}$, emphasizing it as a product of stakeholder engagement. As with any model, the closer $\mathcal{M}$ is to the groundtruth, the better and more reliable is $\mathcal{D}^{CF}$ and, in turn, the results from CST. This problem is often viewed in terms of missing confounders (see, e.g., Kilbertus et al. (2019)), such that what is missing are variables according to a theoretical model or domain expert. Given our focus on defining CST, we take a practical approach to the potential biasedness of the SCM $\mathcal{M}$, viewing it as evidence that is required from and that requires agreement among the stakeholders involved. $\mathcal{D}$ delimits the worldview of the discrimination context. Missing information, such as a confounder, should be addressed by the stakeholders when drawing $\mathcal{M}$ given $\mathcal{D}$.

Coming back to the first theme, the idea of groundtruth for describing human behavior is an open discussion within the fair ML community (see, e.g., Hu and Kohler-Hausmann (2020) and Kasirzadeh and Smart (2021)). We agree that to speak of groundtruth in discrimination testing can be misleading, but that does not diminish the usefulness of using the SCM $\mathcal{M}$ for answering counterfactual questions. As long as the SCM $\mathcal{M}$ is agreed upon by the stakeholders, we argue that the question on whether it amounts to groundtruth or not is secondary.[15] That said, even if the stakeholders agree on a SCM $\mathcal{M}$, it is still possible for other SCMs to describe $\mathcal{D}$. In this work, by always considering a single SCM $\mathcal{M}$, we implicitly work with faithful auxiliary causal knowledge (Peters et al., 2017). What happens when, e.g., the stakeholders agree on or are open to multiple SCMs? We would find ourselves in a competing worlds setting for CST in which the sensitivity of the results

15. For instance, consider that there is already a similar and still unresolved discussion on what it means for individuals to be measurably similar between each other (Westen, 1982).

depend on the faithfulness of the SCM $\mathcal{M}$ (Russell et al., 2017). Future work should explore this line of work as a possible extension to CST.

A non-discriminatory ADM. This work positions CST within indirect discrimination testing for the ADM $b(X) = \hat{Y}$. Recall that we express this type of discrimination using the DAG $A \rightarrow X \rightarrow Y$. Such DAG $\mathcal{G}$ implies the causal relation $X \leftarrow f(A, U)$ in the corresponding SCM $\mathcal{M}$. We use this set up throughout Section 5. As emphasized in the previous two themes, $\mathcal{M}$ and $\mathcal{G}$ describe the dataset $\mathcal{D}$ on which $b()$ is used upon and condition the results of CST. Given this setup, it is reasonable to ask: when does CST detect individual discrimination? Broadly, the answer is when A influences X enough for $b()$ to make decisions that would have been different based on X under a different A. It follows, thus, that CST tests for whether the ADM relies or not on non-neutral information in $\mathcal{D}$. We stress, however, that detecting discrimination is not granted under CST only by having a biased $\mathcal{D}$. It depends, for instance, on how much weight $b()$ gives to elements of X and how these same elements relate to A. Still, since an unbiased $\mathcal{D}$ for high-stake decision-making context is rare to find (Álvarez et al., 2024), it is also reasonable to further ask: how often will CST detect individual discrimination or, put differently, what amounts to a non-discriminatory $b()$ given this setup? The answer to this follow up question is multidisciplinary and more complex.

When testing for indirect discrimination, we essentially suspect the presence of systematic biases—in the form of $A \rightarrow X$—that hinder $b()$ through X and we test wether we are wrong given $\mathcal{D}$. For high-stake settings, like those studied in Section 5, we acknowledge that testing for indirect discrimination using CST sounds like a self-fulfilling prophecy. However, two points are important here. First, even if we do detect indirect discrimination in a context known to be biased, the evidence is one aspect of the discrimination testing pipeline (Weerts et al., 2023). The decision-maker still can justify the use of X in $b()$ as a business requirement. We must keep in mind that CST detects *prima facie* evidence. Second, as argued by legal scholars (Hacker, 2018; Wachter et al., 2020; Xenidis, 2020), the role of indirect non-discrimination law is to address this kind of setting. Methods like CST, at a minimum, raise awareness around the use of X by $b()$ and motivate policies around the business requirements of the decision-maker.

Raising awareness is an important byproduct of discrimination testing methods. As shown in Section 5.2, CST detects the lack of neutrality of X and its impact on $b()$ better than ST. In that example, the bank could successfully justify to a court its use of annual income and account balance despite the results. Further, beyond the business requirements, arguably, there could be a shared societal interest for the bank to make informed decisions, meaning we might want for the bank's $b()$ to be fair but not at the expense of the bank becoming insolvent or applicants going bankrupt (D'Amour et al., 2020; Kozodoi et al., 2022; Schwöbel & Remmers, 2022). What matters, regardless, is showing through CST that the bank uses a non-neutral X, even if it does not translate into validating the complainant's discrimination claim. It helps to acknowledge, through evidence, that we are not in the best possible circumstances as a society. This ties back to the calls made by legal scholars on the role of indirect non-discrimination law and its aim for correcting the present non-neutral status quo. See, e.g., the discussion by Wachter et al. (2020) on substantive equality over formal equality in indirect non-discrimination law. It also ties with calls within fair ML

to acknowledge the present non-neutral status quo as a starting point and view the fair ML tools as corrective measures for achieving a better one. See, e.g., work on revisiting the fairness-accuracy trade-off (Wick et al., 2019) and on non-ideal over ideal fairness (Sahlgren, 2024). Future work should formalize CST and its use of counterfactuals for envisioning and testing for a desired status quo.

7. Conclusion

In this work, we presented counterfactual situation testing (CST), a new actionable and meaningful framework for detecting individual discrimination in a dataset of classifier decisions. We studied both single and multidimensional discrimination, focusing on the indirect setting. For the latter kind, we compared its multiple and intersectional forms and provided the first evidence for the need to recognize intersectional discrimination as separate from multiple discrimination under non-discrimination law. Compared to other methods, such as situation testing (ST) and counterfactual fairness (CF), CST uncovered more cases even when the classifier was counterfactually fair and after accounting for statistical significance. For CF, in particular, we showed how CST equips it with confidence intervals, extending how we understand the robustness of this popular causal fairness definition.

The decision-making settings tackled in this work are intended to showcase the CST framework and, importantly, to illustrate why it is necessary to draw a distinction between idealized and fairness given the difference comparisons when testing for individual discrimination. We hope the results motivate the adoption of the *mutatis mutandis* manipulation over the *ceteris paribus* manipulation. We are aware that the experimental setting could be pushed further by considering higher dimensions or more complex causal structures. We leave this for future work. Further, extensions of CST should consider the impact of using different distance functions for measuring individual similarity (Wilson & Martinez, 1997), and should explore a purely data-driven setup in which the running parameters and auxiliary causal knowledge are derived from the dataset (Cohen, 2013; Peters et al., 2017). Furthermore, extensions of CST should study settings in which the protected attribute goes beyond the binary, such as a high-cardinality categorical or an ordinal protected attribute (Cerda & Varoquaux, 2022). The setting in which the protected attribute is continuous is also of interest, though, in that case we could discretize it (García et al., 2013) and treat it as binary (the current setting) or as a high-cardinality categorical attribute.

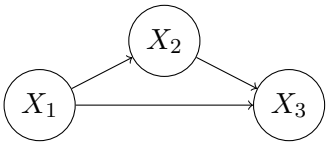
Multidimensional discrimination testing is largely understudied (Roy et al., 2023; Wang et al., 2022). We have set a foundation for exploring the tension between multiple and intersectional discrimination, but future work should further study the problem of dealing with multiple protected attributes and their intersection. It is of interest, for instance, formalizing the case in which one protected attribute dominates the others and the case in which the impact of each protected attribute varies based on individual characteristics. While interaction terms and heterogeneous effects are understudied within SCM, both topics enjoy a well established literature in fields like economics (Wooldridge, 2015), which should enable future work. We hope these extensions and, overall, the fairness given the difference powering the CST framework motivate new work on algorithmic discrimination testing.

Acknowledgments

A preliminary version of this work appeared in Álvarez and Ruggieri (2023). This work has received funding from the European Union's Horizon 2020 research and innovation program under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias"; and from M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI". 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. We thank the three anonymous reviewers for their comments, which helped us improve considerably this work.

Appendix A. Working Example for Generating Counterfactuals

In this section, we present a working example to illustrate counterfactual generation. Given the assumptions we make for a SCM $\mathcal{M}$ (1) in Section 3 and the additional assumption of an additive noise model (ANM) in Section 5, such procedure is straightforward. Suppose we have the following SCM $\mathcal{M}$ and corresponding DAG $\mathcal{G}$:

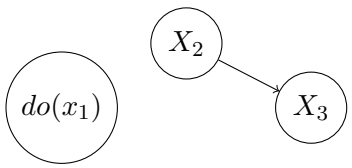


$$\mathcal{M} \begin{cases} X_1 & \leftarrow U_1 \\ X_2 & \leftarrow \alpha \cdot X_1 + U_2 \\ X_3 & \leftarrow \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + U_3 \end{cases}$$

where U_1, U_2, U_3 represent the latent variables, X_1, X_2, X_3 the observed variables, and α, β_1, β_2 the coefficient for the causal effect of, respectively, $X_1 \to X_2$, $X_1 \to X_3$, and $X_2 \to X_3$. Suppose we want to generate the counterfactual for X_3, i.e., X_3^{CF}, had X_1 been equal to x_1. In the **abduction step**, we estimate U_1, U_2, and U_3 given the evidence, or what is observed, under the specified structural equations:

$$\begin{aligned} \hat{U}_1 &= X_1 \\ \hat{U}_2 &= X_2 - \alpha \cdot X_1 \\ \hat{U}_3 &= X_3 - \beta_1 \cdot X_1 + \beta_2 \cdot X_2 \end{aligned}$$

We generalize this step for (1) as $U_j = X_j - f_j(X_{pa(j)})\ \forall X_j \in X$. This step is an individual-level statement on the residual variation under SCM $\mathcal{M}$. It accounts for all that our assignment functions f_j, which are at the population level, cannot explain: i.e., the *error terms*. In the **action step**, we intervene X_1 and set all of its instances equal to x_1 via $do(X_1 := x_1)$ and obtain the intervened DAG $\mathcal{G}'$ and SCM $\mathcal{M}'$:



$$\mathcal{M}' \begin{cases} X_1 & = x_1 \\ X_2 & \leftarrow \alpha \cdot x_1 + U_2 \\ X_3 & \leftarrow \beta_1 \cdot x_1 + \beta_2 \cdot X_2 + U_3 \end{cases}$$

Algorithm 1: CST w/o $(\mathcal{D}, k, \tau, d)$

Input : $\mathcal{D}$ - dataset, k - neighborhood size, τ accepted deviation, d distance function
Output: $\mathcal{R}$ - set of pairs (c, Δ) of protected instance indexes and their $\Delta > \tau$

1 $\mathcal{D}_c = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D} : a_i = 1\}$; // control search space
2 $\mathcal{D}_t = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D} : a_i = 0\}$ // test search space
3 $\mathcal{R} = \emptyset$
4 **for** $(x_c, a_c, \widehat{y}_c) \in \mathcal{D}_c$ **do**
5 $\quad k\text{-}ctr = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D}_c : rank_d(x_c, x_i) \leq k\}$ // control group
6 $\quad k\text{-}tst = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D}_t : rank_d(x_c^{CF}, x_i) \leq k\}$ // test group
7 $\quad p_c = |\{(x_i, a_i, \widehat{y}_i) \in k\text{-}ctr : \hat{y}_i = 0\}|/k$ // fraction of negative decisions for control
8 $\quad p_t = |\{(x_i, a_i, \widehat{y}_i) \in k\text{-}tst : \hat{y}_i = 0\}|/k$ // fraction of negative decisions for test
9 $\quad \Delta = p_c - p_t$ // delta
10 $\quad$ **if** $\Delta > \tau$ **then**
11 $\quad\quad \mathcal{R} = \mathcal{R} \cup \{(c, \Delta)\}$ // add pair to the result
12 **return** $\mathcal{R}$

where no edges come out from X_1 as it has been fixed to x_1. Finally, in the **prediction step**, we combine these two steps to calculate X_3^{CF} under $\hat{U}$ and $\mathcal{M}'$:

$$\begin{aligned} X_3^{CF} &\leftarrow \beta_1 \cdot x_1 + \beta_2 \cdot X_2 + \hat{U}_3 \\ &\leftarrow \beta_1 \cdot x_1 + \beta_2 \cdot (\alpha \cdot x_1 + \hat{U}_2) + \hat{U}_3 \end{aligned}$$

which is done for all instances in X_3. The same three steps apply to X_2 and X_1.

We view the above approach as *frequentist*, in particular, with regard to the abduction step. A more *Bayesian* approach is what is done by Kusner et al. (2017) in which they use a Monte Carlo Markov Chain (MCMC) to draw $\hat{U}$ by updating its prior distribution with the evidence X. In Section 5, we used both approaches and found no difference in the results. In this work, we only present the results for the "frequentist approach" as it is less computationally expensive than running a MCMC.

Appendix B. Supplementary Material

In this section, we present the supplementary material.

B.1 Algorithms

Algorithm 1 reports the pseudo-code of the k-NN CST w/o algorithm. The pseudo-code is self-explanatory. After selecting the control and test search space (lines 1–2) as stated in Definition 4.1, the algorithm iterates over the protected instances. For each of such instances, i.e., the complainant c, it builds (lines 5–6) the control and test groups as the k-nearest neighborhood instances relative to the distance d for x_c and for its counterfactual x_c^{CF} respectively, as stated in (6) and (7). Then, the fractions p_c and p_t of negative decisions for the two groups are computed (lines 7–8) as stated in (9), as well as their difference Δ (line 9). If such a Δ is larger than the accepted deviation τ, then the complainant c and its Δ are added (lines 10–11) to the result $\mathcal{R}$. The pseudo-code of k-NN CST w/ is

a simple variant of Algorithm 1, which adds the search centers x_c and x_c^{CF} into the control and test groups, respectively, and divides by $k+1$ instead of k at lines 7–8.

B.2 Positive Discrimination

Based on the discussion in Section 4.3, we revisit Definitions 4.4 and 4.5 for testing individual positive discrimination under the k-NN CST. We still consider Δp (8) and the converse of the one-sided CI (10).

Definition B.1 (Positive Individual Discrimination)**.** There is (potential) positive individual discrimination in favor of the complainant c if $\Delta p < \tau$, meaning the negative decision outcomes rate for the control group is smaller than for the test group given some accepted deviation $\tau \in [-1, 1]$.

Definition B.2 (Confidence on the Positive Individual Discrimination Claim)**.** A detected (potential) positive discrimination claim for the complainant c by Definition 4.4 is statistically significant with significance level α if the CI $(-\infty, \Delta p + w_\alpha]$ excludes τ.

The concept of positive discrimination also applies to the case of multidimensional discrimination in Section 4.5. In that case, we would re-visit Definitions 4.6 and 4.7 by looking at the opposite effect for Δp across the protected attributes, be it each one of them or their intersection. We do not proceed with redefining these two definitions as we do not showcase them. Again, especially for multidimensional discrimination, our focus is on discrimination *against* protected groups. Further, it is unclear what the legal scholarship views as positive multidimensional discrimination: from Crenshaw (1989) to Xenidis (2020), the focus has been always on traditional discrimination.

Appendix C. Additional Experiments

In this section, we present additional experiments relative to the setup of Section 5.

C.1 Single Positive Discrimination

For the same setup as in Section 5.1 and the same data (factual and counterfactual) as in Section 5.2, we test for positive discrimination using Definitions B.1 and B.2. Regarding counterfactual fairness (CF), we define as positive CF discrimination when the factual has a positive decision outcome, $\hat{y}_c = 1$ but its counterfactual a negative one, $\hat{y}_c^{CF} = 0$. Table 15 summarizes the results for the protected attribute gender.

Unlike all other single discrimination testing results in Section 5, Table 15 shows a ST that detects more cases than CST and a CST and CF that detect no cases at all. Both patterns hold when considering statistical significance. These results are to be expected given how we generated the synthetic data for the loan application scenario. As described in Figure 1, we introduced a negative systematic bias against female applicants in $\mathcal{D}$. This would explain why all the methods based on counterfactual generation—CF, CST w/, and CST w/o—detect zero cases: the generated male counterfactuals in $\mathcal{D}^{CF}$ can only improve over their female factual complainants in $\mathcal{D}$. Hence, $\Delta p < \tau$ is very unlikely to occur when using CST or CF. In other words, in a setting in which the protected individuals are always negatively affected in a systematic way, we should not expect to detect positive

Table 15: Number (and % females) of positive individual discrimination cases in Section 5.2 based on gender. Marked by * are the statistically significant cases

Method	$k = 15$	$k = 30$	$k = 50$	$k = 100$	$k = 250$
CST w/o	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)
ST	45 (2.6%) 41* (2.4%)	50 (2.9%) 48* (2.8%)	77 (4.5%) 55* (3.2%)	118 (6.9%) 93* (5.4%)	159 (9.3%) 120 (7.0%)
CST w/	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)
CF	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)	0 (0.0%) 0* (0.0%)

discrimination under methods that operationalize fairness given the difference. Such results, for the purpose of this work, support our choice to consider only traditional discrimination as it is the most prevalent and important kind of discrimination when we suspect a negative systematic bias against the complainants.

Table 15 raises questions on what kind of comparison is better suited for testing positive discrimination. The fact that ST, which uses an idealized comparison by implementing the CP manipulation, detects discrimination cases in a known biased setting for female applicants puts further into question the role of standard methods like it. Is the idealized comparison suitable for positive discrimination or does Kohler-Hausmann (2018)'s criticism also apply to this setting? Based on these preliminary results, we would argue that the tension between *ceteris paribus* and *mutatis mutandis* manipulations applies also to testing positive discrimination. The stark difference between ST and CST in Table 15 reinforces our view as, essentially, once we account for fairness given the difference in a known biased setting, it is difficult to argue that such a thing as positive discrimination occurs at all. Perhaps this is why this kind of discrimination is not discussed as much by legal scholars looking at indirect discrimination. We plan to revisit these results in future work.

C.2 Single Discrimination Testing

We re-run Section 5.2 and 5.3 for $\tau = 0.05$, keeping all other parameters equal. The results align with the ones we present in the main body. We focus on individual discrimination for all cases, not distinguishing between statistically and non-statistically significant cases.

Table 16 shows the same pattern between the CST versions relative to ST and CF as in Table 1. It illustrates the robustness of our framework. Two points we want to raise regarding Table 16. First, CF, as expected, detects the same number of cases as it always looks for the strict equality between the factual and counterfactual quantities. Second, under $\tau = 0.05$, CST w/ and CST w/o align in the number of cases for larger k sizes. This shows how influential τ can be for detecting discrimination, but also shows that either CST version can tackle the discrimination problem. Tables 17 and 18 show similar results as in, which are the $\tau = 0.05$ counterparts of Tables 10 and 11. The results are expected given

the setup. For both experiments the number of cases drops under $\tau = 0.05$ as we have increased the difficulty of proving the individual discrimination claims.

Table 16: Number and (% w.r.t. females) of cases based on A in Figure 1.

Method	$k = 15$	$k = 30$	$k = 50$	$k = 100$
CST w/o	288 (16.8%)	307 (17.9%)	331 (19.3%)	360 (21.0%)
ST	55 (3.2%)	60 (3.5%)	75 (4.4%)	79 (4.6%)
CST w/	420 (24.5%)	309 (18.1%)	334 (19.5%)	363 (21.2%)
CF	376 (22%)	376 (22%)	376 (22%)	376 (22%)

Table 17: Number (and % w.r.t. non-whites) of cases based on R using Figure 6.

Method	$k = 15$	$k = 30$	$k = 50$	$k = 100$
CST w/o	256 (7.30%)	301 (8.59%)	323 (9.21%)	376 (10.72%)
ST	33 (0.94%)	48 (1.37%)	57 (1.63%)	46 (1.31%)
CST w/	286 (8.16%)	301 (8.59%)	323 (9.21%)	376 (10.72%)
CF	231 (6.59%)	231 (6.59%)	231 (6.59%)	231 (6.59%)

Table 18: Number (and % w.r.t. females) of cases based on G Figure 6.

Method	$k = 15$	$k = 30$	$k = 50$	$k = 100$
CST w/o	78 (0.82%)	105 (1.10%)	224 (2.35%)	231 (2.42%)
ST	77 (0.81%)	92 (0.96%)	181 (1.90%)	185 (1.94%)
CST w/	99 (1.04%)	105 (1.10%)	224 (2.35%)	231 (2.42%)
CF	56 (0.59%)	56 (0.59%)	56 (0.59%)	56 (0.59%)

References

Adams-Prassl, J., Binns, R., & Kelly-Lyth, A. (2023). Directly discriminatory algorithms. *The Modern Law Review*, *86*(1), 144–175.

Adler, J. S. (2019). *Murder in New Orleans: The creation of Jim Crow policing*. University of Chicago Press.

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., & Saha, D. (2018). Automated test generation to detect individual discrimination in AI models. *CoRR*, *abs/1809.03260*.

Álvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Lobo, P. R., Russo, M., Scott, K. M., State, L., Zhao, X., & Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in AI. *Ethics Inf. Technol.*, *26*(2), 31.

Álvarez, J. M., & Ruggieri, S. (2023). Counterfactual situation testing: Uncovering discrimination under fairness given the difference. *EAAMO*, 2:1–2:11.

Álvarez, J. M., & Ruggieri, S. (2024). Mutatis mutandis: Revisiting the comparator in discrimination testing. *CoRR*, *abs/2405.13693*.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics*. Princeton University Press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

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

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

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. *Handbook of Economic Field Experiments*, *1*, 309–393.

Bertrand, M., & Mullainathan, S. (2004). 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.

Black, E., Yeom, S., & Fredrikson, M. (2020). Fliptest: Fairness testing via optimal transport. *FAT**, 111–121.

Bonilla-Silva, E. (1997). Rethinking racism: Toward a structural interpretation. *American Sociological Review*, 465–480.

Bothmann, L., Dandl, S., & Schomaker, M. (2023). Causal fair machine learning via rank-preserving interventional distributions. *AEQUITAS@ECAI*, *3523*.

Carey, A. N., & Wu, X. (2022). The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers Big Data*, *5*, 892837.

Cerda, P., & Varoquaux, G. (2022). Encoding high-cardinality string categorical variables. *IEEE Trans. Knowl. Data Eng.*, *34*(3), 1164–1176.

Chen, Z., Zhang, J. M., Hort, M., Harman, M., & Sarro, F. (2024). Fairness testing: A comprehensive survey and analysis of trends. *ACM Trans. Softw. Eng. Methodol.*, *33*(5), 137:1–137:59.

Chiappa, S. (2019). Path-specific counterfactual fairness. *AAAI*, 7801–7808.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020). Fair regression with Wasserstein barycenters. *NeurIPS*.

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

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, *1989*, 139–167.

Criado-Perez, C. (2019). *Invisible women*. Vintage.

D'Amour, A. (2019). On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *CoRR*, *abs/1902.10286*.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., & Halpern, Y. (2020). Fairness is not static: Deeper understanding of long term fairness via simulation studies. *FAT**, 525–534.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). Fairness through awareness. *ITCS*, 214–226.

EU-FRA. (2018). Handbook on European non-discrimination law [Downloaded in 2023.].

Fix, M., & Struyk, R. J. (1993). *Clear and convincing evidence: Measurement of discrimination in america*. Urban Institute Press.

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

Galhotra, S., Brun, Y., & Meliou, A. (2017). Fairness testing: Testing software for discrimination. *ESEC/SIGSOFT FSE*, 498–510.

García, S., Luengo, J., Sáez, J. A., López, V., & Herrera, F. (2013). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Trans. Knowl. Data Eng.*, *25*(4), 734–750.

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

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871.

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

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. *FAT**, 501–512.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *NIPS*, 3315–3323.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

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

Heikkila, M. (2022). Dutch scandal serves as a warning for europe over risks of using algorithms. *POLITICO*.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *NIPS*, 689–696.

Hu, L., & Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? *FAT**, 513.

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *Int. Conf. on Computer, Control and Communication*, 1–6.

Karimi, A., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. *FAccT*, 353–362.

Karimi, A., von Kügelgen, B. J., Schölkopf, B., & Valera, I. (2020). Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. *NeurIPS*.

Kasirzadeh, A., & Smart, A. (2021). The use and misuse of counterfactuals in ethical machine learning. *FAccT*, 228–236.

Kilbertus, N., Ball, P. J., Kusner, M. J., Weller, A., & Silva, R. (2019). The sensitivity of counterfactual fairness to unmeasured confounding. *UAI*, *115*, 616–626.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *NIPS*, 656–666.

Kleinberg, J. M., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the age of algorithms. *CoRR*, *abs/1902.03731*.

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

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *Eur. J. Oper. Res.*, *297*(3), 1083–1094.

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. *NIPS*, 4066–4076.

Lippert-Rasmussen, K. (2006). The badness of discrimination. *Ethical Theory and Moral Practice*, *9*, 167–185.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D. A., Zemel, R. S., & Welling, M. (2017). Causal effect inference with deep latent-variable models. *NIPS*, 6446–6456.

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2024). When causality meets fairness: A survey. *J. Log. Algebraic Methods Program.*, *141*, 101000.

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

McCandless, L. C., Gustafson, P., & Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, *26*(11), 2331–2347.

Mulligan, D. (2022). Invited talk: Fairness and privacy [At the NeurIPS 2022 Workshop on Algorithmic Fairness through the Lens of Causality and Privacy.].

Nachbar, T. B. (2021). Algorithmic fairness, algorithmic discrimination. *Florida State University Law Review*, *48*, 50.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, *17*(8), 873–890.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd). Cambridge University Press.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *KDD*, 560–568.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.

Plecko, D., & Bareinboim, E. (2024). Causal fairness analysis: A causal toolkit for fair machine learning. *Found. Trends Mach. Learn.*, *17*(3), 304–589.

Plečko, D., & Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, *21*(242), 1–44.

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

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.*, *29*(5), 582–638.

Rooth, D.-O. (2021). Correspondence testing studies. *IZA World of Labor*, *58*.

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

Rose, E. K. (2022). A Constructivist Perspective on Empirical Discrimination Research. *Journal of Economic Literature*, *61*, 906–923.

Rothstein, R. (2017). *The color of law: A forgotten history of how our government segregated america*. Liveright Publishing.

Roy, A., Horstmann, J., & Ntoutsi, E. (2023). Multi-dimensional discrimination in law and machine learning - A comparative overview. *FAccT*, 89–100.

Ruggieri, S., Álvarez, J. M., Pugnana, A., State, L., & Turini, F. (2023). Can we trust fair-AI? *AAAI*, 15421–15430.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. *ACM Trans. Knowl. Discov. Data*, *4*(2), 9:1–9:40.

Russell, C., Kusner, M. J., Loftus, J. R., & Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. *NIPS*, 6414–6423.

Sahlgren, O. (2024). What's impossible about algorithmic fairness? *Philosophy and Technology*, *37*(4).

Schneider, E. C. (2008). *Smack: Heroin and the American city*. University of Pennsylvania Press.

Schwöbel, P., & Remmers, P. (2022). The long arc of fairness: Formalisations and ethical discourse. *FAccT*, 2179–2188.

Sen, M., & Wasow, O. (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, *19*(1), 499–522.

Thanh, B. L., Ruggieri, S., & Turini, F. (2011). K-NN as an implementation of situation testing for discrimination discovery and prevention. *KDD*, 502–510.

Totenberg, N. (2023). Supreme court guts affirmative action, effectively ending race conscious admissions. *NPR*. https://www.npr.org/2023/06/29/1181138066/affirmative-action-supreme-court-decision

Tschantz, M. C. (2022). What is proxy discrimination? *FAccT*, 1993–2003.

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

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, *123*, 735.

Wang, A., Ramaswamy, V. V., & Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *FAccT*, 336–349.

Weerts, H. J. P., Xenidis, R., Tarissan, F., Olsen, H. P., & Pechenizkiy, M. (2023). Algorithmic unfairness through the lens of EU non-discrimination law: Or why the law is not a decision tree. *FAccT*, 805–816.

Westen, P. (1982). The empty idea of equality. *Harvard Law Review*, 537–596.

Wick, M. L., Panda, S., & Tristan, J. (2019). Unlocking fairness: A trade-off revisited. *NeurIPS*, 8780–8789.

Wightman, L. F. (1998). *LSAC national longitudinal bar passage study*. Law School Admission Council.

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *J. Artif. Intell. Res.*, *6*, 1–34.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage Learning.

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

Yang, K., Loftus, J. R., & Stoyanovich, J. (2021). Causal intersectionality and fair ranking. *FORC*, *192*, 7:1–7:20.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *ICML (3)*, *28*, 325–333.

Zhang, L., Wu, Y., & Wu, X. (2016). Situation testing-based discrimination discovery: A causal inference approach. *IJCAI*, 2718–2724.