Consistent End-to-End Estimation for Counterfactual Fairness

Yuchen Ma* YUCHEN.MA@LMU.DE
Valentyn Melnychuk* MELNYCHUK@LMU.DE
Dennis Frauen
Stefan Feuerriegel
LMU Munich & Munich Center for Machine Learning (MCML), Munich, Germany
**Equal contribution.*

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Counterfactual fairness seeks to ensure that the prediction for an individual is the same as that in a counterfactual world under a different sensitive attribute. However, achieving counterfactual fairness in practice is challenging due to several issues. In this paper, we discuss two key issues and suggest workarounds for them. (1) Counterfactual fairness involves counterfactual quantities, which rely on unobservable outcomes and inference and which require additional, non-trivial counterfactual identifiability assumptions. Here, we explicitly state the *counterfactual identifiability* assumptions necessary to guarantee point identification of counterfactuals up to a measure-preserving indeterminacy (MPI). Without such counterfactual identifiability assumptions, counterfactual fairness can not be *reliably* fulfilled, even with infinite data. (2) Even assuming counterfactual identifiability holds, many existing baselines for counterfactual fairness lack consistency in estimating counterfactuals and thus may yield predictions that are unfair. In our work, we redefine the notion of consistency so that it (i) is compatible with the counterfactual identifiability up to an MPI, and (ii) allows to enforce worst-case counterfactual fairness. Following this definition of consistency, we develop an end-to-end method that allows for *consistent estimation* of counterfactuals (which we call a generative counterfactual fairness network). Once the counterfactuals are consistently estimated, we can use regularization as an in-processing approach to enforce the worst-case counterfactual fairness in prediction models. With our work, we aim to spur a discussion of when and where counterfactual fairness can be reliably achieved to ensure a safe and ethical use.

Keywords: counterfactual fairness, counterfactual inference, causal fairness, identifiability

1. Introduction

In this paper, we focus on the notion of ***counterfactual fairness*** (Kusner et al., 2017). The notion of counterfactual fairness has recently received significant attention (e.g., (Kusner et al., 2017; Garg et al., 2019; Xu et al., 2019; Kim et al., 2021; Abroshan et al., 2022; Garg et al., 2019; Grari et al., 2023; Ma et al., 2023; Zuo et al., 2023)). Formally, counterfactual fairness ensures that the prediction for an individual is the same as that in a counterfactual world under a different sensitive attribute.

However, achieving counterfactual fairness is challenging for two main reasons. (1) *Counterfactual identifiability:* Counterfactual fairness involves counterfactual quantities, namely, counterfactual distributions. They are based on unobservable counterfactual outcomes and are located at level 3 of Pearl's causality ladder.[1] Hence, inferring the counterfactual distributions reliably from observational

1. See Pearl (2009) for Pearl's causality ladder.

data requires additional, non-trivial counterfactual identifiability assumptions.[2] Furthermore, the counterfactual identification results strictly guide the design of downstream estimation methods for counterfactual fairness. ② *Consistent estimation:* Even when the counterfactual identifiability assumptions are satisfied, the consistent estimation of counterfactual quantities from observational data poses further challenges (Xia et al., 2023). Existing methods often fail to consistently estimate the counterfactual distributions and, thus, may yield predictions that are unfair.

Motivated by the above challenges, we now turn to existing methods that study counterfactual fairness and evaluate them in light of the two key issues described above (see Table 1). Kusner et al. (2017) first proposed a conceptual algorithm for counterfactual fairness, and the conceptual algorithm was later extended to neural methods (Pfohl et al., 2019; Kim et al., 2021; Grari et al., 2023). However, none of these works has thoroughly studied the interplay between counterfactual identifiability and consistent estimation (see Appendix B). Our aim is to fill this gap by clarifying the theoretical requirements for counterfactual identifiability and proposing ways to ensure consistent estimation. In doing so, we aim to broaden our understanding of when counterfactual fairness can be *reliably* achieved – and when it might not, potentially leading to harmful downstream consequences.

Why should we care about counterfactual identifiability and consistent estimation? Both are necessary to reliably learn the "true" counterfactual distribution (i.e., how an individual's outcome would change if only a sensitive attribute were altered) (Manski, 2009). Without both identifiability *and* consistency, no amount of data can reliably uncover the "true" counterfactual relationship (Manski, 2009), rendering any fairness claims ungrounded. Even if counterfactual identifiability holds, inconsistent estimation can systematically bias estimated counterfactual distributions and lead to unfair predictions.

Crucially, we argue that the assumptions that we state later for counterfactual identifiability should not be seen as weaknesses but rather as boundary conditions that clarify when a method can be used reliably – and when it cannot. Methods that fail to state such assumptions operate without any assurance and potentially produce unreliable or even harmful outcomes. In contrast, with counterfactual identifiability (and a downstream method for consistent estimation), we gain theoretical guarantees that counterfactual fairness assessments reflect the underlying causal reality. Hence, regardless of whether these assumptions seem strict or not for real-world applications, our main goal is to clarify what assumptions may be necessary to *reliably* achieve counterfactual fairness in practice.

In this paper, *we theoretically study the consistent, end-to-end estimation of the counterfactual distributions for counterfactual fairness.*[3] We thus aim to address the above issues and suggest workarounds for them. ① *Counterfactual identifiability:* We explicitly discuss identifiability assumptions necessary to guarantee point identification of counterfactuals up to a measure-preserving indeterminacy (MPI) (Xi and Bloem-Reddy, 2023). Without such counterfactual identifiability assumptions, counterfactual fairness can not be reliably fulfilled, even with infinite data (Melnychuk et al., 2023). While some of the counterfactual identifiability concepts are known in causal inference, we adapt them to counterfactual fairness. ② *Consistent estimation:* To allow for consistent estimation, we redefine the notion of consistency so that it (i) is compatible with the counterfactual identifiability

2. Identifiability is defined as an ability to infer the true values of a model's underlying parameters after obtaining an infinite number of observations from it (Manski, 2009). Importantly, identification is *different* from estimability because methods that act as heuristics may return estimates, but they do not correspond to the true value (D'Amour, 2019).
3. Code is available at https://github.com/yccm/consistent-estimation-gcfn.

Table 1: Key neural methods for counterfactual fairness prediction (see Appendix B for details).

	Counterfactual identifiability	Counterfactual estimation (→ issue 2)		
	(→ issue 1)	Method	End-to-end	Consistency
(Pfohl et al., 2019)	✗	mCEVAE	✗ (AAP)	✗
(Kim et al., 2021)	✗	DCEVAE	✗ (AAP)	✗
(Grari et al., 2023)	✗	ADVAE	✗ (AAP)	✗
Our work	✓ (POF + BGMs)	GCFN	✓	✓ (MPI consistency)

Legend: POF + BGMs (standard potential outcomes framework assumptions + bijective generative models); AAP (abduction-action-prediction); MPI (measure-preserving indeterminacy).

up to an MPI (Xi and Bloem-Reddy, 2023), and (ii) allows to enforce worst-case counterfactual fairness. Following this definition of consistency, we develop an end-to-end method that allows for *consistent estimation* of counterfactual distributions, which we later call *Generative Counterfactual Fairness Network* (GCFN). Once the counterfactuals are consistently estimated, we can use them in a regularization term for any predictive model. In this way, we implement an in-processing approach to enforce the worst-case counterfactual fairness in prediction models.

Ethical discussion. (1) We emphasize that the assumptions we introduce should *not* be viewed as limitations but rather as explicit boundary conditions clarifying exactly when *any* method can reliably achieve counterfactual fairness. Hence, our primary purpose is to spur a discussion around the limitations of the counterfactual fairness notion and demonstrate, both theoretically and empirically, how existing methods can fail. In that spirit, one of the implications of our work is that more caution is needed as to whether counterfactual fairness can be reliably achieved in practice. Here, our method may serve as an *auditing tool* for practitioners to detect fairness violations in datasets – and thus uncover biases in decision-making practice.

2. Problem setup

Notation: Capital letters such as V, X, A, M denote random variables and small letters v, x, a, m denote their realizations from corresponding domains $\mathcal{V}, \mathcal{X}, \mathcal{A}, \mathcal{M}$. Further, $\mathbb{P}(M)$ is the probability distribution of M; $\mathbb{P}(M \mid A = a)$ is a conditional distribution (level 1); $\mathbb{P}(M_a)$ the interventional (potential) distribution on M when setting A to a (level 2); and $\mathbb{P}(M_{a'} \mid A = a, M = m)$ the counterfactual distribution of M had A been set to a' given evidence $A = a$ and $M = m$ (level 3).

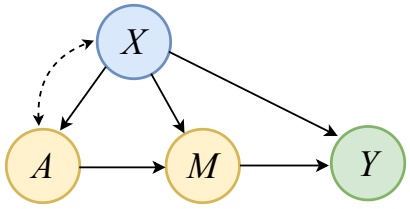

Figure 1: Causal graph. The nodes represent: sensitive attribute, A; covariate, X; mediator, M; target, Y. $\rightarrow$ represents direct causal effect; $\leftarrow\!\!\cdots\!\!\rightarrow$ represents potential presence of hidden confounders.[4]

We follow the *Standard Fairness Model* (Plečko and Bareinboim, 2024), shown in Fig. 1. (1) We consider a sensitive attribute $A \in \mathcal{A}$. (2) We consider predictors $V = M \cup X$, which we grouped into two sets with different roles: mediators $M \in \mathcal{M}$, which are possibly caused by the sensitive attribute; and covariates $X \in \mathcal{X}$, which are not caused by the sensitive attribute. (3) We also consider a target $Y \in \mathcal{Y}$. See Appendix C for more details. In our work, A can be a categorical variable with multiple categories k and X and M can be multi-dimensional. For ease of notation, we use $k = 2$,

i.e., $\mathcal{A} = \{0, 1\}$, to present our method below. We later present an extension to settings where the sensitive attribute has multiple categories (see Appendix F).

Splitting predictors into mediators and covariates: For our theoretical derivations, we categorize the predictors V into covariates X (pre-treatment; i.e., independent of the sensitive attribute) and mediators M (post-treatment; i.e., direct descendants of the sensitive attribute). This step is crucial for achieving counterfactual identifiability later. Although it may appear to be a constraint, it in fact clarifies the roles of different predictors and later helps us make the underlying assumptions explicit. Of note, our method requires only knowing whether a variable is affected by the sensitive attribute, rather than specifying the entire causal structure among these mediators (Kim et al., 2021).

In practice, identifying mediators is typically straightforward – especially in domains where sensitive attributes (e.g., gender) clearly influence other variables (e.g., income). Such knowledge is common in practice (Schröder et al., 2023). We discuss practical strategies for this in Appendix C. If practitioners are unsure, a "worst-case" approach is to include all predictors as mediators (rather than as covariates).

Example: In a typical loan application setting, the sensitive attribute A might be gender, and the final outcome Y is loan approval. The mediators M (e.g., income, education, or employment history) are downstream variables that gender can influence (e.g., because gender norms may affect opportunities and thus shape education or employment paths). By contrast, covariates X are the non-mediator or pre-treatment variables (e.g., age, location), which are not affected by the sensitive attribute.

Estimand: Here, we use the terminology of potential outcomes framework (Rubin, 1974). Under our causal graph, the dependence of M on A implies that interventions on the sensitive attribute A also mean potential changes in the mediator M. We use subscripts such as M_a to denote the potential outcome of M when intervening on $A = a$. Similarly, Y_a denotes the potential outcome of Y. Furthermore, for $k = 2$, A is the factual, and A' is the counterfactual value of the sensitive attribute.

We are interested in the notion of *counterfactual fairness* (Kusner et al., 2017). Therein, the aim is to learn the prediction of a target Y to be *counterfactual fair* with respect to some given sensitive attribute A. Let $h(X, M) = \hat{Y}$ denote the predicted target from some prediction model, which only depends on covariates and mediators. Formally, for $k = 2$ w.l.o.g., h achieves counterfactual fairness if under any context $X = x$, $A = a$, and $M = m$, the following counterfactual distributions are equal:

$$\mathbb{P}\left(h(x, M_a) \mid X = x, A = a, M = m\right) = \mathbb{P}\left(h(x, M_{a'}) \mid X = x, A = a, M = m\right). \tag{1}$$

This definition illustrates the need to care about the counterfactual mediator distributions.

3. Counterfactual identifiability assumptions

We now discuss the assumptions *necessary* assumptions to *reliably* estimate causal quantities from observational data, namely, the counterfactual mediator distributions, and therefore to enforce counterfactual fairness.

4. The dashed line allows for a correlation between X and A in our framework. Note that, if there is no dashed edge between X and A, it is actually a stronger assumption, because it forbids the edge between X and A to have any hidden confounders. However, our setting is more general and allows for the existence of confounders.

Potential outcomes framework: We adopt the standard assumptions of the *potential outcomes framework* (Rubin, 1974). **(1)** *Consistency:* The observed mediator is identical to the potential mediator given a certain sensitive attribute. Formally, for each unit of observation, $A = a \Rightarrow M = M_a$. **(2)** *Overlap:* For all x such that $\mathbb{P}(X = x) > 0$, we have $0 < \mathbb{P}(A = a \mid X = x) < 1, \forall a \in \mathcal{A}$. **(3)** *Unconfoundedness:* Conditional on covariates X, the potential outcome M_a is independent of sensitive attribute A, i.e. $M_a \perp\!\!\!\perp A \mid X$.

Under the consistency assumption (1), the left side of Eq. (1) simplifies to the delta (point mass) distribution:

$$\mathbb{P}\left(h(x, M_a) \mid X = x, A = a, M = m\right) = \delta\left(h(x, m)\right). \tag{2}$$

Also, the assumptions (1)–(3) help to identify interventional (level 2) distributions (e.g., $\mathbb{P}(h(x, M_{a'}) \mid X = x)$) or even some counterfactual (level 3) distributions (e.g., $\mathbb{P}\left(h(x, M_{a'}) \mid X = x, A = a\right)$ when $k = 2$). However, to point identify the counterfactual mediator distribution under the full context, as required in the right side of Eq. (1), additional assumptions are needed (Melnychuk et al., 2023).

Bijective generative mechanisms: We further focus on bijective generation mechanisms (BGMs) (Nasr-Esfahany et al., 2023; Melnychuk et al., 2023). **(4)** *BGM assumption:* Let the observational distribution $\mathbb{P}_{X,A,M} = \mathbb{P}_{\mathrm{f}}$ be induced by an SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$ with

$$\begin{aligned} &\mathbf{V} = \{X, A, M, Y\}, \quad \mathbf{U} = \{U_{XA}, U_M, U_Y\}, \\ &\mathcal{F} = \{f_X(u_{XA}), f_A(x, u_{XA}), f_M(x, a, u_M), f_Y(x, m, u_Y)\}, \quad \mathbb{P}(\mathbf{U}) = \mathbb{P}(U_{XA})\mathbb{P}(U_M)\mathbb{P}(U_Y), \end{aligned} \tag{3}$$

and with the causal graph as in Figure 1. Let $\mathcal{M} \subseteq \mathbb{R}$. Then, f_M is called a bijective generation mechanism (BGM) if f_M is a strictly increasing (decreasing)[5] continuously-differentiable wrt. u_M.

Given the BGM assumption in (4), the counterfactual distribution of the mediator simplifies to one of two possible point mass distributions:

$$\mathbb{P}(h(x, M_{a'}) \mid X = x, A = a, M = m) = \delta\big(h(x, \mathbb{F}^{-1}(\pm\mathbb{F}(m; x, a) \mp 0.5 + 0.5; x, a'))\big), \tag{4}$$

where $\mathbb{F}(\cdot; x, a)$ and $\mathbb{F}^{-1}(\cdot; x, a)$ are a CDF and an inverse CDF of $\mathbb{P}(M \mid X = x, A = a)$, respectively, and $\delta(\cdot)$ is a Dirac-delta distribution. For the derivation of Eq. (4), we refer to Lemma B.2 in (Nasr-Esfahany et al., 2023) and to Corollary 3 in (Melnychuk et al., 2023). Furthermore, as Eq. (4) yields two point mass distributions, we say that the counterfactual distribution is *identifiable up to a measure-preserving indeterminacy* (MPI) (Xi and Bloem-Reddy, 2023).

Remark 1 *The difference between the two solutions in Eq. (4) is negligibly small when the conditional standard deviation of the mediator is small (see Appendix D for the derivation).*

Remark 2 *The BGM assumption can be naturally extended to high-dimensional mediators when the dimensionality of U_M matches the dimensionality of M and f_M is bijective wrt. to u_M. Yet, in this case, there is a continuum of solutions in the class of continuously differentiable functions (Chen and Gopinath, 2000).*

5. Importantly, under mild conditions (if the conditional density of the mediator has finite values), this result holds in the more general class of BGMs with non-monotonous continuously differentiable functions when U_M is one dimensional (Melnychuk et al., 2023).

Discussion: The BGM assumption includes many popular identifiable SCMs as special cases, such as ANM (Peters et al., 2014), LSNM (Immer et al., 2023), and PNL (Zhang and Hyvärinen, 2010). Thus, this assumption can be seen as one of the most general assumptions that lead to counterfactual point identifiability. Notwithstanding, ***any*** method (and not just ours) aimed at counterfactual fairness would need to fulfill the BGM assumption to ensure counterfactual identifiability and consistent estimation.

Nevertheless, it is hard to argue whether real-world datasets usually satisfy the BGM assumption. As discovered by Melnychuk et al. (2023), the relaxation of the BGM assumption not only immediately leads to counterfactual point non-identifiability but also to non-informative partial counterfactual identification bounds. Still, it can be intuitively re-formulated (Melnychuk et al., 2023) as follows: In f_M, the sensitive attribute A is assumed to interact only with the observed covariates X and not with the exogenous noise U_M. Many real-world data-generation mechanisms and phenomena, if studied closely, can be said to satisfy this assumption (e.g., simulators in physics and medicine, but also neuroscience and behavioral processes). We offer a more extensive discussion in Appendix D.1.

We acknowledge that the counterfactual identifiability is an existing result in causal inference (e.g., Lemma B.2 in Nasr-Esfahany et al. (2023) and Corollary 3 in Melnychuk et al. (2023)). However, surprisingly, the exact assumptions needed for counterfactual identifiability – and hence for the downstream consistent estimation – in the context of counterfactual fairness are often overlooked in previous literature (e.g., (Pfohl et al., 2019; Kim et al., 2021; Grari et al., 2023)). We thus believe that stating the assumptions explicitly will help clarify when counterfactual fairness can be reliably achieved, because the results also guide the design of a consistent estimation strategy.

4. End-to-end consistent estimation

Overview: Here, we introduce a *consistent*, end-to-end estimation method called Generative Counterfactual Fairness Network (GCFN). An overview of our method is in Fig. 2. GCFN proceeds in two steps: **Step 1** learns the counterfactual distribution of the mediator. **Step 2** uses the generated counterfactual mediators in an in-processing approach (Plečko and Bareinboim, 2024) to enforce the worst-case counterfactual fairness in prediction models. For the latter, we introduce a *counterfactual mediator regularization*. The pseudocode can be found in Appendix E.

Why do we need counterfactuals of the mediator? Different from existing methods for causal effect estimation (Bica et al., 2020; Yoon et al., 2018), we are *not* interested in obtaining potential outcomes for the target Y ($\widehat{=}$ level 2 in Pearl's causality ladder). Instead, we are interested in the counterfactuals for the mediator M, which then captures the entire influence of the sensitive attribute and its descendants on the target ($\widehat{=}$ level 3). Thus, by training the prediction model with our *counterfactual mediator regularization*, we remove the information from the sensitive attribute to ensure fairness while keeping the rest of the useful information in the data to maintain predictive power.

4.1. Step 1: Generating counterfactuals

In Step 1, we aim to generate counterfactuals of the mediator (since the ground-truth counterfactual mediator is unavailable). A key strength is that Step 1 *directly* learns transformations of factual mediators to counterfactuals without the intermediate step of inferring latent variables (as in existing works). As a result, we eliminate the need for the abduction-action-prediction procedure (Pearl, 2009)

and avoid the complexities and potential inaccuracies of inferring and then using latent variables for prediction.

Formally, a deterministic generator G produces the counterfactual of the mediators given observational data, while a discriminator D differentiates the factual mediator from the generated counterfactual mediators. This adversarial training process encourages G to learn the counterfactual distribution of the mediator. We further control the quality of the counterfactual through the adversarial loss: the deterministic generator seeks to generate counterfactual mediators in a way that minimizes the probability that the discriminator can differentiate between factual mediators and counterfactual mediators, while the discriminator seeks to maximize the probability of correctly identifying the factual mediator. We replace the generated factual mediator with the observed factual mediator before passing it as input to the discriminator (this is also known as teacher forcing).

How do we prevent Step 1 from reproducing factuals (and rather learn counterfactuals)? (1) *Intuitively*, this is due to the adversarial loss: the input of the discriminator D contains the counterfactual and the factual, but *their location is randomized in a mini-batch*. Through this randomization, we ensure that the counterfactuals are learned, as otherwise the discriminator would obtain a perfect score. (2) *Theoretically*, we prove that the generator converges to the true counterfactual distribution $\mathbb{P}(M_{a'} \mid X = x, A = a, M = m)$, see Proposition 1 later. (3) *Empirically*, we show that our generated counterfactual mediators diverge considerably from their factual counterparts while aligning well with ground-truth counterfactuals (see Appendix J). Together, this explains why Step 1 can successfully learn the true counterfactuals.

Counterfactual deterministic generator G: The deterministic generator G learns the counterfactual distribution of the mediator, i.e., $\mathbb{P}\left(M_{a'} \mid X = x, A = a, M = m\right)$. Formally, $G : \mathcal{X} \times \mathcal{A} \times \mathcal{M} \to \mathcal{M}$. G takes the factual sensitive attribute A, the factual mediator M, and the covariates X as inputs, sampled from the joint (observational) distribution $\mathbb{P}_{X,A,M}$, denoted as $\mathbb{P}_{\mathrm{f}}$ for short. G outputs two potential mediators, $\hat{M}_0$ and $\hat{M}_1$, from which one is factual and the other is counterfactual. For notation, we use $G\left(X, A, M\right)$ to refer to the output of the deterministic generator. Thus, we have

$$G\left(X, A, M\right)_a = \hat{M}_a \quad \text{for} \quad a \in \{0, 1\}. \tag{5}$$

In our deterministic generator G, we intentionally output not only the counterfactual mediator but also the factual mediator, even though the latter is observable. The reason is that we can use it to further stabilize the training of the deterministic generator. For this, we introduce a reconstructive loss $\mathcal{L}_{\mathrm{f}}$, which we use to ensure that the generated factual mediator $\hat{M}_A$ is similar to the observed factual mediator M. Formally, we define the reconstruction loss with an L_2-norm $\|\cdot\|_2$:

$$\mathcal{L}_{\mathrm{f}}(G) = \mathbb{E}_{(X,A,M)\sim\mathbb{P}_{\mathrm{f}}}\left[\left\|M - G\left(X, A, M\right)_A\right\|_2^2\right]. \tag{6}$$

Counterfactual discriminator D: The discriminator D is carefully adapted to our setting. In an ideal world, we would have D discriminate between real vs. fake counterfactual mediators; however, the counterfactual mediators are not observable. Instead, we train D to discriminate between factual mediators vs. generated counterfactual mediators. Note that this is different from the conventional discriminators in GANs that seek to discriminate real vs. fake samples (Goodfellow et al., 2014). Formally, our discriminator D is designed to differentiate the factual mediator M (as observed in the data) from the generated counterfactual mediator $\hat{M}_{A'}$ (as generated by G).

We modify the output of G before passing it as input to D: We replace the generated factual mediator $\hat{M}_A$ with the observed factual mediator M. We denote the new, combined data by $\tilde{G}\left(X, A, M\right)$,

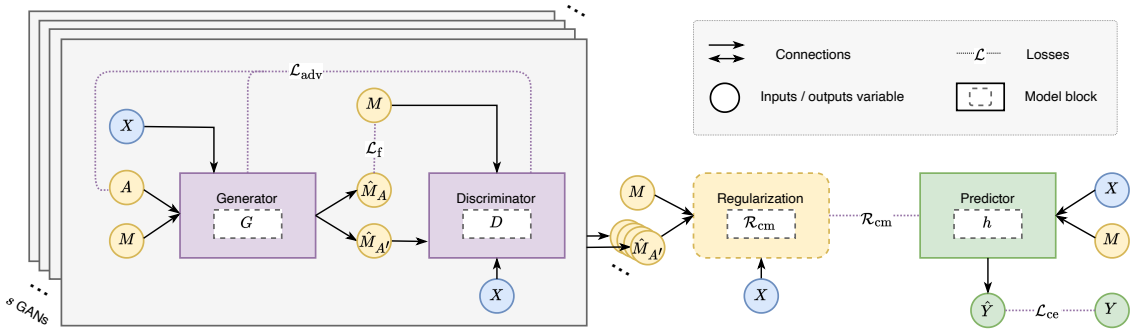

Figure 2: **Overview of our GCFN for consistent estimation.** *Step 1:* A deterministic generator G takes (X, A, M) as input and outputs $\hat{M}_A$ and $\hat{M}_{A'}$. A discriminator D then differentiates the observed factual mediator M from the generated counterfactual mediator $\hat{M}_{A'}$. We train s different generator-discriminator pairs and consider the worst-case counterfactual fairness. *Step 2:* We then use generated counterfactual mediator $\hat{M}_{A'}$ in our counterfactual mediator regularization $\mathcal{R}_{\text{cm}}$. We take the supremum of $\mathcal{R}_{\text{cm}}$ to choose the most 'unfair' generator. Therefore, we enforce the worst-case counterfactual fairness for the prediction model $h(x, m)$.

which is defined via

$$\tilde{G}\left(X, A, M\right)_a = \begin{cases} M, & \text{if } A = a, \\ G\left(X, A, M\right)_a, & \text{if } A = a'. \end{cases} \tag{7}$$

The discriminator D then determines which component of $\tilde{G}$ is the observed factual mediator and thus outputs the corresponding probability. Formally, for the input $(X, \tilde{G})$, the output of the discriminator D is

$$D(X, \tilde{G})_a = \hat{\mathbb{P}}(M = \tilde{G}_a \mid X, \tilde{G}) = \hat{\mathbb{P}}(A = a \mid X, \tilde{G}). \tag{8}$$

Adversarial training: Step 1 is trained in an adversarial manner: (i) the deterministic generator G seeks to generate counterfactual mediators in a way that minimizes the probability that the discriminator can differentiate between factual mediators and counterfactual mediators, while (ii) the discriminator D seeks to maximize the probability of correctly identifying the factual mediator. We thus use an adversarial loss $\mathcal{L}_{\text{adv}}$ given by

$$\mathcal{L}_{\text{adv}}(G, D) = \mathbb{E}_{(X,A,M)\sim\mathbb{P}_{\text{f}}}\left[\log\left(D(X, \tilde{G}\left(X, A, M\right))_A\right)\right]. \tag{9}$$

Overall, our Step 1 is trained through an adversarial training procedure with a minimax problem as

$$\min_G \max_D \mathcal{L}_{\text{adv}}(G, D) + \alpha\mathcal{L}_{\text{f}}(G), \tag{10}$$

with a hyperparameter α on $\mathcal{L}_{\text{f}}$.

Then, under mild counterfactual identifiability conditions, the counterfactual distribution of the mediator, i.e., $\mathbb{P}\left(M_{a'} \mid X = x, A = a, M = m\right)$, is *consistently estimated* by Step 1 (up to a measure-preserving indeterminacy (Xi and Bloem-Reddy, 2023)). We later state this formally in Proposition 1.

4.2. Step 2: Counterfactual fairness with counterfactual mediator regularization

In Step 2, we use the output of Step 1 to train a prediction model under counterfactual fairness in a supervised way. For this, we introduce our *counterfactual mediator regularization* that enforces counterfactual fairness w.r.t the sensitive attribute. Let h denote our prediction model (e.g., a neural network). We define our *counterfactual mediator regularization* $\mathcal{R}_{\text{cm}}(h, G)$ as

$$\mathcal{R}_{\text{cm}}(h, G) = \mathbb{E}_{(X,A,M)\sim\mathbb{P}_{\text{f}}}\left[\left\|h\left(X, M\right) - h\left(X, \hat{M}_{A'}\right)\right\|_2^2\right], \tag{11}$$

where $\hat{M}_{A'} = G(X, A, M)_{A'}$. Our counterfactual mediator regularization has three important characteristics: (1) It is non-trivial. Different from traditional regularization, our $\mathcal{R}_{\text{cm}}$ is not based on the representation of the prediction model h, but it *involves a counterfactual quantity $\hat{M}_{A'}$ that is otherwise not observable*. (2) Our $\mathcal{R}_{\text{cm}}$ is not used to constrain the learned representation (e.g., to avoid overfitting), but it is used to change the actual learning objective to achieve the property of counterfactual fairness. (3) Our $\mathcal{R}_{\text{cm}}$ fulfills theoretical properties. Specifically, we show later that, under some conditions, our regularization actually optimizes against counterfactual fairness and thus should learn our task as desired.

The overall loss $\mathcal{L}(h)$ is as follows. We fit the prediction model h using a cross-entropy loss $\mathcal{L}_{\text{ce}}(h)$. We further integrate the above counterfactual mediator regularization $\mathcal{R}_{\text{cm}}(h, G)$ into our overall loss $\mathcal{L}(h)$. For this, we introduce a weight $\lambda \geq 0$ to balance the trade-off between prediction performance and the level of counterfactual fairness. Formally, we have

$$\mathcal{L}(h) = \mathcal{L}_{\text{ce}}(h) + \lambda \sup_{G\in\mathcal{G}} \mathcal{R}_{\text{cm}}(h, G), \tag{12}$$

where $\mathcal{G}$ is a set of all the deterministic generators minimizing Eq. 10, and the supremum over this set chooses the most 'unfair' generator. A large value of λ increases the weight of $\mathcal{R}_{\text{cm}}$, thus leading to a prediction model that is strict with regard to counterfactual fairness, while a lower value allows the prediction model to focus more on producing accurate predictions. As such, λ offers additional flexibility to decision-makers as they tailor the prediction model based on the fairness needs in practice.

The supremum in Eq. (12) chooses the most 'unfair' generator. Therefore, we enforce a *worst-case counterfactual fairness*, as the ground-truth counterfactual distribution is only identifiable up to an MPI (see Sec. 3).

Computational efficiency: In the above method, we choose the supremum over s generator-discriminator pairs to rigorously bound the worst-case performance and to address the MPI. In practice, it is often sufficient to rely on a single generator-discriminator pair, which considerably reduces the computational overhead while maintaining strong empirical performance (Appendix K).

4.3. Theoretical results

Proposition 1 states that our deterministic generator consistently estimates the counterfactual distribution of the mediator $\mathbb{P}(M_{a'} \mid X = x, A = a, M = m)$ up to a measure-preserving indeterminacy (MPI).

Proposition 1 (Consistent estimation of the counterfactual distribution (up to an MPI)) *If the deterministic generator in Step 1 is a continuously differentiable function with respect to M, then it consistently estimates one of the identified counterfactual distributions of the mediator (Eq. 4).*

Proof Intuitively, we first prove that, given an optimal discriminator, the deterministic generator of Step 1 estimates the distribution of potential mediators for counterfactual sensitive attributes. We then prove that the outputs of the deterministic generator, conditional on the factual mediator $M = m$, estimate $\mathbb{P}(M_{a'} \mid X = x, A = a, M = m)$. Details are in Appendix D. ■

Corollary 3 (MPI identifiability of high-dimensional mediators) *The results of the Proposition 1 generalize to multi-dimensional mediators. In this case, the deterministic generator consistently estimates one of the implicitly given counterfactual distributions:*

$$\mathbb{P}(M_{a'} = m' \mid X = x, A = a) = \int_{\mathcal{M}} \delta(G(x, a, m)_{a'} - m')\, \mathbb{P}(M = m \mid X = x, A = a)\, \mathrm{d}m. \tag{13}$$

Next, we provide a theoretical analysis to show that our proposed counterfactual mediator regularization is effective in ensuring counterfactual fairness for predictions. Following Grari et al. (2023), we measure the level of counterfactual fairness CF via $\mathbb{E}[\|(h\,(X, M) - h\,(X, M_{A'}))\|_2^2]$. It is straightforward to see that the smaller CF is, the more counterfactual fairness the prediction model achieves.

We show that by empirically measuring our generated counterfactual of the mediator, we can thus quantify to what extent counterfactual fairness CF is fulfilled in the prediction model. We give an upper bound in the following proposition.

Proposition 2 (Counterfactual mediator regularization bound[6]) *Given the prediction model h that is Lipschitz continuous with a Lipschitz constant $\mathcal{C}$, we have*

$$\mathbb{E}\left[\|(h(X, M) - h(X, M_{A'})\|_2^2\right] \leq \mathcal{C}\, \mathbb{E}\left[\left\|M_{A'} - \hat{M}_{A'}\right\|_2^2\right] + \sup_{G \in \mathcal{G}} \mathcal{R}_{\mathrm{cm}}(h, G), \text{ for every } G \in \mathcal{G}, \tag{14}$$

where $\hat{M}_{A'} = G(X, A, M)_{A'}$ and $\mathcal{G}$ is a set of all deterministic generators, minimizing the Eq. 10.

Proof See Appendix D. ■

The inequality in Proposition 2 states that the influence from the sensitive attribute on the target variable is upper-bounded by (i) the estimation of counterfactual mediators (first term) and (ii) the counterfactual mediator regularization (second term). (i) The first term does not depend on h, and, given Proposition 1, reduces to zero as there exists a deterministic generator in $\mathcal{G}$, which consistently estimates counterfactuals. Hence, by reducing (ii) the second term $\mathcal{R}_{\mathrm{cm}}$ for all the deterministic generators through minimizing our training loss in Eq. 12, we can effectively enforce the predictor to be more counterfactual fair.[7]

7. Details how we ensure Lipschitz continuity in h are in Appendix H.

5. Experiments

In our experiments, *we primarily aim to confirm the theoretical guarantees of our approach by demonstrating how enforcing counterfactual identifiability assumptions and ensuring consistent estimation translate into practical gains*.

Baselines: We compare our method against the following state-of-the-art approaches: (1) **CFAN** (Kusner et al., 2017); (2) **CFUA** (Kusner et al., 2017); (3) **mCEVAE** (Pfohl et al., 2019); (4) **DCE-VAE** (Kim et al., 2021); (5) **ADVAE** (Grari et al., 2023); (6) **HSCIC** (Quinzan et al., 2024); and (7) **CFGAN** (Xu et al., 2019). Details are in Appendix H, including hyperparameter tuning.

Performance metrics: Methods for counterfactual fairness aim at both: (i) achieve high accuracy while (ii) ensuring counterfactual fairness, which essentially yields a multi-criteria decision-making problem. To this end, we follow standard procedures and reformulate the multi-criteria decision-making problem using a utility function $U_\gamma(accuracy, CF) : \mathbb{R}^2 \mapsto \mathbb{R}$, where CF is the metric for measuring counterfactual fairness from Sec. 4.2. We define the utility function as $U_\gamma(accuracy, CF) = accuracy - \gamma \times CF$ with a given utility weight γ. A larger utility U_γ is better. The weight γ depends on the application and is set by the decision-maker; here, we report results for a wide range of weights $\gamma \in \{0.1, \ldots, 1.0\}$.

5.1. Results for (semi-)synthetic data

We explicitly focus on (semi-)synthetic data, which allows us to compute the true counterfactuals and thus validate the effectiveness of our method. We follow previous works that simulate a fully synthetic dataset for performance evaluations (Kim et al., 2021; Quinzan et al., 2024). Detailed results are in Appendix I.1

Setting: The Law School (LSAC) dataset (Wightman, 1998) contains information about the law school admission records. We use the LSAC dataset to construct semi-synthetic datasets with two different data-generating mechanisms. (See Appendix G). We predict whether a candidate passes the bar exam and where *gender* is the sensitive attribute. We simulate 101,570 samples and use 20% as the test set.

Results: Results are shown in Fig. 3. We make the following findings. Our GCFN performs best. Compared to the baselines, the performance gain from our GCFN is large (up to $\sim$30%). The performance gain for our GCFN tends to become larger for larger γ. The strong performance of our GCFN in the semi-synthetic dataset with "sin" function demonstrates that our tailored method can capture complex counterfactual distributions. Together, this confirms that ensuring counterfactual identifiability and consistent estimation can lead to better performance.

Empirical validation that we learn counterfactuals (and reproduce factuals): Intuitively, one may think that Step 1 would simply copy the factual mediators because the counterfactual mediators can not be observed during training. However, this is ***not*** the case due to the adversarial training process of the deterministic generator. Contrarily, we successfully learn the counterfactual mediators. Experiment results and details are in Appendix J. We observe that the generated counterfactual mediator is similar to the ground-truth counterfactual mediator, while the factual and the generated counterfactual mediators are highly dissimilar. In other words, our model can correctly learn the counterfactual mediators.

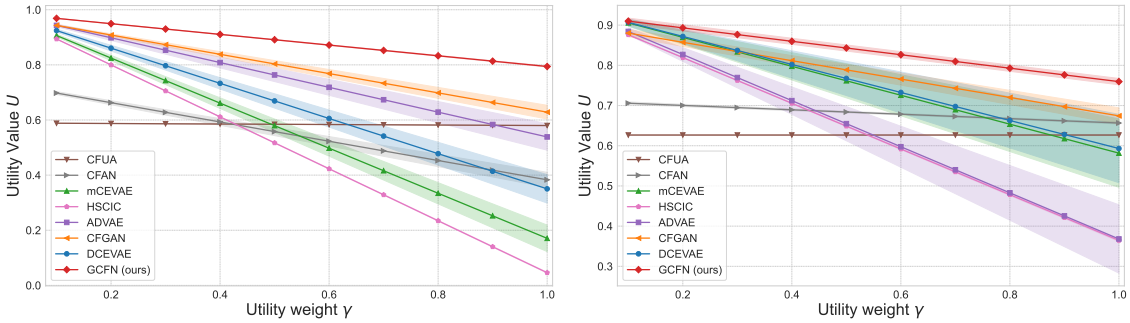

Figure 3: **Results for LSAC dataset with two different data-generating mechanisms.** A larger utility is better. Shown: mean $\pm$ std over 5 runs.

5.2. Results for real-world datasets

We demonstrate the applicability of our method to real-world data using **UCI Adult** (Asuncion and Newman, 2007) and **COMPAS** (Angwin et al., 2016). The results are in Appendix L.

6. Related work

Fairness notions for predictions: Over the past years, the machine learning community has developed an extensive series of fairness notions for predictive tasks so that one can train unbiased machine learning models (see Appendix B for a detailed overview). In this paper, we focus on *counterfactual fairness* (Kusner et al., 2017), because it is closely related to legal terminology (Barocas and Selbst, 2016; De Arteaga et al., 2022).

Predictions under counterfactual fairness: Originally, Kusner et al. (2017) introduced a conceptual algorithm to achieve predictions under counterfactual fairness. The idea is to first infer a set of latent background variables and subsequently train a prediction model using these inferred latent variables and non-descendants of sensitive attributes. However, the conceptual algorithm requires knowledge of the ground-truth structural causal model, which makes it impractical.

State-of-the-art approaches build upon the above conceptual algorithm but integrate neural learning techniques, typically by using VAEs. To approximate the counterfactuals, these approaches build upon an abduction-action-prediction procedure (Pearl, 2009), yet which involves latent variables and may introduce inaccuracies, including wrong results due to a lack of identifiability of the latent variables and consistent estimation (D'Amour, 2019). Prominent examples are mCEVAE (Pfohl et al., 2019), DCEVAE (Kim et al., 2021), and ADVAE (Grari et al., 2023). In general, these methods build upon the abduction-action-prediction procedure (Pearl, 2009): first compute the posterior distribution of the latent variables, given the observational data and a prior on latent variables. Based on that, they compute the implied counterfactual distributions, which can either be utilized directly for predictive purposes or can act as a constraint incorporated within the training loss. Further details are in Appendix B. In sum, the methods in Pfohl et al. (2019); Kim et al. (2021) and Grari et al. (2023) are our main baselines.

More important and at the core of our contribution are the following two **shortcomings** (see Table 1): (1) The existing methods proceed with inferring latent variables ***without*** establishing proper counterfactual identifiability. Therefore, there is no guarantee that the inferred latent variables match the ground truth underlying distribution (D'Amour, 2019). (2): The existing methods do ***not*** allow for

consistent estimation. For example, under the BGM assumption, the counterfactual distributions are represented by point mass distributions, which renders VAEs a wrong modeling choice. Motivated by these shortcomings, we aim to develop a consistent estimation method compatible with the counterfactual identification results.

7. Discussion

Flexibility: In this paper, our main aim was to construct an end-to-end method to *consistently estimate* counterfactual fairness and to confirm the theoretical properties through experiments. Our method is highly flexible and works with both discrete and categorical sensitive attributes (see Appendix F).

Limitations: Our results also point to challenges as to whether counterfactual fairness is identifiable in practice. This is not inherent to our method but points towards a more fundamental challenge of this fairness notion. Hence, we hope to spur discussions about the potential ethical risks of using counterfactual fairness.

Acknowledgments

This paper is supported by the DAAD program "Konrad Zuse Schools of Excellence in Artificial Intelligence", sponsored by the Federal Ministry of Education and Research.

References

Mahed Abroshan, Mohammad Mahdi Khalili, and Andrew Elliott. Counterfactual fairness in synthetic data generation. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. In *ProPublica*, 2016.

Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: The works of Judea Pearl*, 2022.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 2016.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *NeurIPS*, 2020.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Conference on Fairness, Accountability, and Transparency*, 2019.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAacT*, 2019.

Scott Chen and Ramesh Gopinath. Gaussianization. In *NeurIPS*, 2000.

Silvia Chiappa. Path-specific counterfactual fairness. In *AAAI*, 2019.

Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. In *AISTATS*, 2019.

Maria De Arteaga, Stefan Feuerriegel, and Maytal Saar-Tsechansky. Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management*, 2022.

Pietro G Di Stefano, James M Hickey, and Vlasios Vasileiou. Counterfactual fairness: Removing direct effects through regularization. In *arXiv preprint*, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, 2012.

Jake Fawkes, Robin Evans, and Dino Sejdinovic. Selection, ignorability and challenges with causal fairness. In *Conference on Causal Learning and Reasoning*. PMLR, 2022.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, 2015.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *AAAI*, pages 219–226, 2019.

Ian Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks, April 2017.

Ian Goodfellow, Jean Pouget.Abadie, Mehdi Mirza, Bing Xu, David Warde.Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Adversarial learning for counterfactual fairness. *Machine Learning*, 2023.

Nina Grgic.Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, 2016.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, 2016.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *ICML*, 2023.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Fair algorithms for infinite and contextual bandits. *arXiv preprint*, 2016.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTATS*, 2020.

Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il.Chul Moon. Counterfactual fairness with disentangled causal effect variational autoencoder. In *AAAI*, 2021.

Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *ICLR*, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NeurIPS*, 2017.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NeurIPS*, 2017.

Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. Learning for counterfactual fairness from observational data. In *KDD*, 2023.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *FAacT*, 2019.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint*, 2020.

Charles F. Manski. *Identification for prediction and decision*. Harvard University Press, Cambridge, MA, 2009.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Partial counterfactual identification of continuous outcomes with a curvature sensitivity model. In *NeurIPS*, 2023.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *AAAI*, 2018.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *ICML*, 2023.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *NeurIPS*, 2020.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 2009.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 2014.

Stephen R. Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H Shah. Counterfactual reasoning for fair clinical risk prediction. In *Machine Learning for Healthcare Conference*, 2019.

Drago Plečko and Elias Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024.

Francesco Quinzan, Cecilia Casolo, Krikamol Muandet, Yucen Luo, and Niki Kilbertus. Learning counterfactually invariant predictors. *TMLR*, 2024. ISSN 2835-8856.

Amirarsalan Rajabi and Ozlem Ozmen Garibay. Tabfairgan: Fair tabular data generation with generative adversarial networks. In *Machine Learning and Knowledge Extraction*, 2022.

Lucas Rosenblatt and R Teal Witter. Counterfactual fairness is basically demographic parity. In *AAAI*, 2023.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 1974.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *International Conference on Management of Data*, 2019.

Maresa Schröder, Dennis Frauen, and Stefan Feuerriegel. Causal fairness under unobserved confounding: A neural sensitivity framework. In *ICLR*, 2023.

Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. DECAF: Generating fair synthetic data using causally-aware generative networks. In *NeurIPS*, 2021.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: An application to recidivism prediction. In *arXiv preprint*, 2018.

Yixin Wang, Dhanya Sridhar, and David Blei. Adjusting machine learning decisions for equal opportunity and counterfactual fairness. *TMLR*, 2023.

Linda F Wightman. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. In *ERIC*, 1998.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *AISTATS*, 2023.

Kevin Muyuan Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *ICLR*, 2023.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. FairGAN: Fairness-aware generative adversarial networks. In *ICBD*, 2018.

Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *IJCAI*, 2019.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AIES*, 2018.

Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Workshop on Causality: Objectives and Assessment at NIPS 2008*, 2010.

Zeyu Zhou, Tianci Liu, Ruqi Bai, Jing Gao, Murat Kocaoglu, and David I Inouye. Counterfactual fairness by combining factual and counterfactual predictions. *NeurIPS*, 2024.

Zhiqun Zuo, Mahdi Khalili, and Xueru Zhang. Counterfactually fair representation. *NeurIPS*, 2023.

Appendix A. Background

Notation: Capital letters such as U denote a random variable and small letters u its realizations from corresponding domains $\mathcal{U}$. Bold capital letters such as $\mathbf{U} = \{U_1, \ldots, U_n\}$ denote finite sets of random variables. Further, $\mathbb{P}(Y)$ is the distribution of a variable Y.

SCM: A structural causal model (SCM) (Pearl, 2009) is a 4-tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$, where $\mathbf{U}$ is a set of exogenous (background) variables that are determined by factors outside the model; $\mathbf{V} = \{V_1, \ldots, V_n\}$ is a set of endogenous (observed) variables that are determined by variables in the model (i.e., by the variables in $\mathbf{V} \cup \mathbf{U}$); $\mathcal{F} = \{f_1, \ldots, f_n\}$ is the set of structural functions determining $\mathbf{V}, v_i \leftarrow f_i\left(\mathrm{pa}\left(v_i\right), u_i\right)$, where $\mathrm{pa}\left(V_i\right) \subseteq \mathbf{V} \backslash V_i$ and $U_i \subseteq \mathbf{U}$ are the functional arguments of f_i; $\mathbb{P}(\mathbf{U})$ is a distribution over the exogenous variables $\mathbf{U}$.

Potential outcome: Let X and Y be two random variables in $\mathbf{V}$ and $\mathbf{u} = \{u_1, \ldots, u_n\} \in \mathcal{U}$ be a realization of exogenous variables. The potential outcome $Y_x(\mathbf{u})$ is defined as the solution for Y of the set of equations $\mathcal{F}_x$ evaluated with $\mathbf{U} = \mathbf{u}$ (Pearl, 2009). That is, after $\mathbf{U}$ is fixed, the evaluation is deterministic. $Y_x(\mathbf{u})$ is the value variable Y would take if (possibly contrary to observed facts) X is set to x, for a specific realization $\mathbf{u}$. In the rest of the paper, we use Y_x as the short for $Y_x(\mathbf{U})$.

Observational distribution: A structural causal model $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$ induces a joint probability distribution $\mathbb{P}(\mathbf{V})$ such that for each $Y \subseteq \mathbf{V}$, $\mathbb{P}^{\mathcal{M}}(Y = y) = \sum_u \mathbb{1}(Y(\mathbf{u}) = y)\mathbb{P}(\mathbf{U} = \mathbf{u})$ where $Y(\mathbf{u})$ is the solution for Y after evaluating $\mathcal{F}$ with $\mathbf{U} = \mathbf{u}$ (Bareinboim et al., 2022).

Counterfactual distributions: A structural causal model $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$ induces a family of joint distributions over counterfactual events $Y_x, \ldots, Z_w$ for any $Y, Z, \ldots, X, W \subseteq \mathbf{V}$: $\mathbb{P}^{\mathcal{M}}\left(Y_x = y, \ldots, Z_w = z\right) = \sum_u \mathbb{1}\left(Y_x(\mathbf{u}) = y, \ldots, Z_w(\mathbf{u}) = z\right) \mathbb{P}(\mathbf{U} = \mathbf{u})$ (Bareinboim et al., 2022). This equation contains variables with different subscripts, which syntactically represent different potential outcomes or counterfactual worlds.

Causal graph: A graph $\mathcal{G}$ is said to be a causal graph of SCM $\mathcal{M}$ (Pearl, 2009; Bareinboim et al., 2022), if represented as a directed acyclic graph (DAG), where each endogenous variable $V_i \in \mathbf{V}$ is a node; there is an edge $V_i \longrightarrow V_j$ if V_i appears as an argument of $f_j \in \mathcal{F}$ ($V_i \in \mathrm{pa}(V_j)$); there is a bidirected edge $V_i \dashleftarrow\!\dashrightarrow V_j$ if the corresponding $U_i, U_j \subset \mathbf{U}$ are correlated ($U_i \cap U_j \neq \varnothing$) or the corresponding functions f_i, f_j share some $U_{ij} \in \mathbf{U}$ as an argument.

Appendix B. Extended related work

B.1. Fairness

Recent literature has extensively explored different fairness notions (e.g., Dwork et al., 2012; Feldman et al., 2015; Grgic.Hlaca et al., 2016; Hardt et al., 2016; Joseph et al., 2016; Zafar et al., 2017; Wadsworth et al., 2018; Madras et al., 2018; Zhang et al., 2018; Pfohl et al., 2019; Salimi et al., 2019; Celis et al., 2019; Chen et al., 2019; Madras et al., 2019; Di Stefano et al., 2020) For a detailed overview, we refer to Makhlouf et al. (2020) and Plečko and Bareinboim (2024). There have also been theoretical advances (e.g., Fawkes et al., 2022; Rosenblatt and Witter, 2023) but these are orthogonal to ours.

Existing fairness notions can be loosely classified into notions for group- and individual-level fairness, as well as causal notions, some aim at path-specific fairness (e.g., Nabi and Shpitser, 2018; Chiappa, 2019). We adopt the definition of counterfactual fairness from Kusner et al. (2017).

Counterfactual fairness (Kusner et al., 2017): Given a predictive problem with fairness considerations, where A, X and Y represent the sensitive attributes, remaining attributes, and output of interest respectively, for a causal model $\mathcal{M} = \langle \mathbf{V} = \{A, X, Y\}, \mathbf{U}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$, prediction model $\hat{Y} = h(X, A, \mathbf{U})$ is counterfactual fair, if under any context $X = x$ and $A = a$,

$$\mathbb{P}\left(\hat{Y}_a(\mathbf{U}) \mid X = x, A = a\right) = \mathbb{P}\left(\hat{Y}_{a'}(\mathbf{U}) \mid X = x, A = a\right), \tag{15}$$

for any value a' attainable by A. This is equivalent to the following formulation:

$$\mathbb{P}\left(h(X_a(\mathbf{U}), a, \mathbf{U}) \mid X = x, A = a\right) = \mathbb{P}\left(h(X_{a'}(\mathbf{U}), a', \mathbf{U}) \mid X = x, A = a\right). \tag{16}$$

Our paper adapts the latter formulation by doing the following. First, we make the prediction model independent of the sensitive attributes A, as they could only make the predictive model unfairer. Second, given the general non-identifiability of the posterior distribution of the exogenous noise (= non-identifiability of latent variables), i.e., $\mathbb{P}\left(\mathbf{U} \mid X = x, A = a\right)$, we consider only the prediction models dependent on the observed covariates. Third, we split observed covariates X into pre-treatment covariates (confounders) and post-treatment covariates (mediators). Thus, we yield our definition of a fair predictor in Eq. 1.

B.2. Kusner's counterfactual fairness

Originally, Kusner et al. (2017) introduced a conceptual algorithm to achieve predictions under counterfactual fairness. The idea is to first infer a set of latent background variables and subsequently train a prediction model using these inferred latent variables and non-descendants of sensitive attributes. Kusner et al. (2017) provided only a conceptual algorithm, while only later works proceeded by offering actual instantiations. In particular, Kusner et al. (2017) did not clarify how to learn latent variables in practice and did not prove the identifiability of the inferred latent variables by a model. As such, the conceptual algorithm does not provide any theoretical guarantees for achieving counterfactual fairness in the final prediction models. Furthermore, the conceptual algorithm is not able to directly learn fairness prediction from a given dataset and a given causal graph, but instead it requires the specification of additional structural equations and thus requires further domain knowledge. Without knowing the ground-truth structural causal model, Kusner et al. (2017) can not have counterfactual identifiability and can not achieve counterfactual fairness. In sum, this makes the conceptual algorithm – and any other instantiation building upon it – impractical.

B.3. Discussion about theoretical guarantees in counterfactual fairness prediction

Current works related to counterfactual fairness prediction often state that their proposed methods satisfy counterfactual fairness under certain conditions (Pfohl et al., 2019; Kim et al., 2021; Grari et al., 2023; Zuo et al., 2023; Wang et al., 2023; Zhou et al., 2024). However, none of these papers consider or discuss the identifiability of counterfactuals. These papers either require some implied strong assumptions about the identifiability of counterfactuals (which they do not specify) or fully ignore the identifiability at all, and hope their method can somehow manage to learn the correct counterfactuals. As we discuss below, this is often not true because of which these methods may learn predictions that are unfair.

Zuo et al. (2023) claims that they form the counterfactual prediction by drawing from the counterfactual distribution. However, it does not clarify how they exactly managed to learn this counterfactual distribution, and neither does the paper prove why the learned counterfactual distribution by this model should be identifiable. They just give a conceptual algorithm, yet it may learn an unfair objective.

Wang et al. (2023); Zhou et al. (2024) involve the step of learning the latent variables and later training a prediction model using these inferred latent variables. These papers do not clarify how they managed to learn the latent variables and do not have any proper counterfactual identification guarantees for learned latent variables. For example, they can use, for example, a VAE to estimate the latent variable but it is not guaranteed that it can be correctly identified, leading to risks that the latent variable is often estimated incorrectly. Again, this leads to predictions that can be actually unfair.

Methods that act as heuristics as those above may return estimates, but these estimates do not correspond to the true value. So, theoretically, these methods can converge against predictions that are not fair but unfair.

B.4. Benefits over latent variable baselines

Importantly, the latent variable baselines for counterfactual fairness (e.g., mCEVAE (Pfohl et al., 2019), DCEVAE (Kim et al., 2021), and ADVAE (Grari et al., 2023)) are far from being easy as they do not rely on off-the-shelf methods. Rather, they also learn a latent variable in non-trivial ways. The inferred latent variable U should be independent of the sensitive attribute A while representing all other useful information from the observation data. However, there are two main challenges: (1) The latent variable U is *not* identifiable. (2) It is very *hard* to learn such U to satisfy the above independence requirement, especially for high-dimensional or other more complicated settings. Hence, we argue that baselines based on some custom latent variables are highly challenging.

Because of (1) and (2), there are **no** theoretical guarantees for the VAE-based methods. Hence, it is mathematically unclear whether they actually learn the correct counterfactual fair predictions. In fact, there is even rich empirical evidence that VAE-based methods are often *suboptimal*. VAE-based methods use the estimated variable U in the first step to learn the counterfactual outcome $\mathbb{P}\left(\hat{Y}_{a'}(\mathbf{U}) \mid X = x, A = a, M = m\right)$. The inferred, non-identifiable latent variable can be correlated with the sensitive attribute, which may harm fairness, or it might not fully represent the rest of the information from data and harm prediction performance.

Importantly, the latent variable baselines do *not* allow for identifiability. In causal inference, "identifiability" refers to a mathematical condition that permits a causal quantity to be measured from observed data (Pearl, 2009). Importantly, identification is *different* from estimation because methods

that act as heuristics may return estimates, but they do not correspond to the true value. For the latter, see D'Amour (2019) where the authors provide several concerns that, if a latent variable is not unique, it is possible to have local minima, which leads to unsafe results in causal inference.

Non-identifiable for VAE-based methods have been shown in prior works of literature. In a recent paper Xia et al. (2023), the authors show that VAE-based counterfactual inference does *not* allow for counterfactual identifiability. The results directly apply to variational inference-based methods, which *do* not have proper counterfactual identification guarantees. Also, the result from non-linear ICA (which is the task of variational autoencoders) shows that the latent variables are *non-identifiable* (Khemakhem et al., 2020). In simple words, VAE-based methods can estimate the latent variable but it is *not* guaranteed that it can be correctly identified, leading to risks that the latent variable is often estimated incorrectly. Note that non-identifiability of the latent variables leads to the counterfactual non-identifiability (Melnychuk et al., 2023). Hence, VAE-based methods can **not** ensure that they correctly learn counterfactual fairness, only our method does so.

B.5. Generating fair synthetic datasets

A different literature stream has used generative models to create fair synthetic datasets (e.g., Xu et al., 2018, 2019; van Breugel et al., 2021; Rajabi and Garibay, 2022). Importantly, the task and objective here are *different* from ours. Here, relevant to us is only one method called CFGAN (Xu et al., 2019). However, it is vastly different from our method in many aspects (see next section for details).

B.6. Difference from CFGAN

Even though CFGAN also employs GANs, it is vastly **different** from our method.

1. *Different tasks:* CFGAN is designed for fair data generation tasks, while our model is designed for learning predictors to be counterfactual fairness. Hence, both address **different tasks**. The training objectives are **different**: CFGAN learns to **mimic factual data**. In our method, the deterministic generator **learns the counterfactual distribution of the mediator** through the discriminator distinguishing factual from counterfactual mediators.

2. *Different architectures:* CFGAN employs **two** generators, each aimed at simulating the original causal model and the interventional model, and two discriminators, which ensure data utility and causal fairness. We only employ a streamlined architecture with a **single** deterministic generator and a discriminator. Further, fairness enters both architectures at **different places**. In CFGAN, fairness is ensured through the GAN setup, whereas our method ensures fairness in a second step through our counterfactual mediator regularization.

3. *Different mathematical objectives:* CFGAN is proposed to synthesize a dataset that satisfies counterfactual fairness in the sampled data. However, a recent paper of Abroshan et al. (2022) has shown that CFGAN is actually considering interventions (=level 2 in Pearl's causality ladder) and **not** counterfactuals (=level 3).[8] Hence, CFGAN does **not** fulfill the counterfactual

8. In the context of Pearl's causal hierarchy**, interventional and counterfactual queries are completely different concepts (Bareinboim et al., 2022). (1) Interventional queries are located on level 2 of Pearl's causality ladder. Interventional queries are of the form $P(y \mid do(x))$. Here, the typical question is "What if? What if I do X?", where the activity is "doing". (2) Counterfactual queries are located on level 3 of Pearl's causality ladder. Counterfactual queries are of the

fairness notion, but a different notion based on the do-operator (intervention). For details, we refer to Abroshan et al. (2022), Definition 5 therein, called "Discrimination avoiding through causal reasoning"): A generator is said to be fair if the following equation holds: for any context $A = a$ and $X = x$, for all value of y and $a' \in \mathcal{A}$, $P(Y = y \mid X = x, do(A = a)) = P(Y = y \mid X = x, do(A = a'))$, which is different from the counterfactual fairness $P\left(\hat{Y}_a = y \mid X = x, A = a\right) = P\left(\hat{Y}_{a'} = y \mid X = x, A = a\right)$..

4. *No theoretical guarantee for CFGAN:* CFGAN **lacks theoretical support** for its methodology (no counterfactual identifiability guarantees or counterfactual fairness level). In contrast, our method strictly satisfies the principles of counterfactual fairness and provides theoretical guarantees on the counterfactual fairness level. In sum, **only our method offers theoretical guarantees** for the task at hand.

5. *Suboptimal performance of CFGAN:* Even though CFGAN can, in principle, be applied to counterfactual fairness prediction, it is **suboptimal**. The reason is the following. Unlike CFGAN, which generates complete synthetic data under causal fairness notions, our method only generates counterfactuals of the mediator as an intermediate step, resulting in minimal information loss and better inference performance than CFGAN. Furthermore, since CFGAN needs to train the dual-generator and dual-discriminator together and optimize two adversarial losses, it is more difficult for stable training, and thus its method is less robust than ours.

In sum, even though CFGAN also employs GANs, it is **vastly different from our method**.

B.7. Deep generative models for estimating causal effects

There are many papers that leverage generative adversarial networks and variational autoencoders to estimate causal effects from observational data (Louizos et al., 2017; Kocaoglu et al., 2018; Yoon et al., 2018; Pawlowski et al., 2020; Bica et al., 2020). We later borrow some ideas of modeling counterfactuals through deep generative models, yet we emphasize that those methods aim at estimating causal effects but *without* fairness considerations.

form $P(y_x \mid x', y')$, where x' and y' are different values that X, Y took before. Here, the typical question is "What if I had acted differently?", where the activity is "imagining" had a different treatment selection been made in the beginning. Hence, the main difference is that the counterfactual of y is conditioned on the post-treatment outcome (factual outcome) of y and a different x (where x takes a different value than x'). For details, we kindly refer to the papers of Bareinboim et al. (2022); Pearl (2009) for a more technical definition of why intervention and counterfactual are two entirely different concepts.

Appendix C. Causal graph

C.1. Selection of the causal graph

We choose a causal graph in a way that we do not assume complete knowledge about the causal graph, just whether a variable is pre- or post-treatment, that is, whether a variable is a descendant of the sensitive attribute. Furthermore, our assumption of the causal graph is aligned with a standard approach in counterfactual fairness, i.e., with the *Standard Fairness Model* in Plečko and Bareinboim (2024)

C.2. Further clarification

In this section, we clarify why our framework is flexible and broadly applicable.

C.2.1. ARROW FROM X TO A

Below, we clarify that the arrow from X to A is not a limitation but a more general setting.

In causal inference, having an arrow means there is *no* additional assumption, while the absence of an arrow means that there is an assumption. So in our causal graph, the presence of an arrow from X to A indicates the allowance of a causal relationship between variables, that is X can be a direct cause of A, this is permissible/allowed in our framework, but the direct causal relationship $X \to A$ is not a necessity. If there is no arrow from X to A, it is actually a stronger assumption, because it forbids X to be the cause of A.

C.3. Practical example of our causal graph

In practice, it is common and typically straightforward to choose which variables act as mediators M through domain knowledge (Nabi and Shpitser, 2018; Kim et al., 2021; Plečko and Bareinboim, 2024). Hence, mediators M are simply all variables that can potentially be influenced by the sensitive attribute. All other variables (except for A and Y) are modeled as covariates X. For example, consider a job application setting where we want to avoid discrimination by gender. Then, A is gender, and Y is the job offer. Mediators are, for instance, education level or work experience, as both are potentially influenced by gender. In contrast, age is a covariate because it is not influenced by gender.

C.3.1. MEDIATOR SELECTION

In this section, we clarify that having the mediator in the causal graph is a standard setting (Schröder et al., 2023). Below, we give a detailed introduction to the mediator selection and further offer a step-by-step process on how practitioners can adapt their settings to our graph.

1. Knowledge of M is required for *any* theoretically grounded method that aims to identify/estimate counterfactual effects. It is well-known in the causal inference literature that knowledge of pre-treatment and post-treatment variables is required to identify a causal effect (even in Layer 2 of Pearl's causal hierarchy) (Pearl, 2009). Hence, such knowledge must be available to identify stronger fairness notions such as counterfactual fairness, which lie on layer 3 of Pearl's hierarchy. Works that do not distinguish between post- and pre-treatment variables have no hope of achieving identifiability on the interventional level, and also no hope

for the (harder) counterfactual level. The assumption is *consistent* with established literature on causal fairness, and the corresponding model is called "the *Standard Fairness Model*" in the literature (Plečko and Bareinboim, 2024). Mediators are part of a standard causal model, particularly in the context of fairness, where sensitive attributes are often involved due to some discrimination-related issues. Mediators naturally occur in practical applications where there are sensitive attributes. Knowledge of M is given in many practical scenarios where our method could be applied.

2. Knowledge of mediators is often realistic as they are usually available mediators in observed data (De Arteaga et al., 2022). The reason is located in the definition of sensitive attributes. Sensitive attributes are of concern to (dis)advantage certain groups of society because the sensitive attribute is known to influence other variables, which may act as proxies that drive some outcome. For example, consider a loan application and take gender as a sensitive attribute. Obviously, gender is sensitive in loan applications as it is known to not only be responsible for discrimination, but it naturally affects many other secondary outcomes (e.g., income), which makes it so important to mitigate discrimination with respect to the sensitive attribute in the first place. If gender were not affecting outcomes broadly, one would not necessarily see the importance of mitigating discrimination with regard to it, and one would thus not consider it a sensitive attribute.

3. Our assumptions regarding M are weaker than in some existing literature: it is important to clarify that we have a more general setting about mediators. Our method does not require complete knowledge of the entire causal graph. In some previous works, such as Kim et al. (2021), detailed knowledge of the causal relationships among mediators is required, while our methodology does not need to specify the causal relationships among them. This makes our framework more flexible and broadly applicable.

4. What if the mediator M is not correctly identified? In this case, one could still employ a "worst-case" approach by including all variables we are unsure about in M, which gives us a heuristic worst-case approach. This is similar to other methods that do not distinguish between X and M (Xu et al., 2019). However, those methods that use such heuristics (including baselines) have no theoretical grounding. In summary, we would like to emphasize that the ability of our method to incorporate knowledge of M is an advantage rather than a disadvantage of our method. If M can be correctly identified, then our method provides strong performance (accuracy) with theoretical guarantees on the identifiability of counterfactual fairness. If not, the worst-case approach is still possible for applying our method.

In summary, it is important to clarify that we have a more *general setting* about mediators. Our method does not require complete knowledge of the entire causal graph. Our primary objective is to identify whether variables are pre- or post-treatment, in other words, determining if they are descendants of the sensitive attribute. Variables potentially affected by the sensitive attribute are classified as mediators. In some previous works, detailed knowledge of the causal relationships among mediators is required, while our methodology does not. This makes our framework *more flexible and broadly applicable.*

Appendix D. Theoretical results

D.1. Discussion about BGM assumption

The bijective generation mechanism (BGM) (Nasr-Esfahany et al., 2023) is required to ensure the counterfactual identifiability of our method. The BGM assumption is crucial for the identification of the counterfactuals. It includes many popular identifiable SCMs as special cases, e.g., ANM (Peters et al., 2014), LSNM and (Immer et al., 2023), and PNL (Zhang and Hyvärinen, 2010). Thus, this assumption can be seen as one of the most general assumptions that lead to counterfactual point identifiability.

It is hard to argue whether real-world datasets usually satisfy the BGM assumption. Rather, this assumption provides a guideline for which datasets it is – in principle – possible to provide answers to the counterfactual questions and for which not. As discovered by Melnychuk et al. (2023), the relaxation of the BGM assumption not only immediately leads to counterfactual point non-identifiability but also to non-informative partial counterfactual identification bounds. Still, it can be intuitively re-formulated (Melnychuk et al., 2023) as follows: In f_M, the sensitive attribute, A, is assumed to interact only with the observed covariates, X, and not with the exogenous noise, U_M. Many real-world data-generation mechanisms/phenomena, if studied closely, can be said to satisfy this assumption (e.g., simulators in physics and medicine, but also neuroscience and behavioral processes).

Current literature on counterfactual fairness does not even have any theoretical guarantees on the counterfactual identifiability of their methods or provide no guarantees of the correctness of counterfactual fairness achieved on either synthetic datasets or real-world datasets. We believe that having some assumptions to make our method with theoretical guarantees is better than methods with no correctness guarantee at all, and thus helps us to understand where and when counterfactual fairness can be fulfilled from a theoretical point of view.

Remark 1 *The difference between the two solutions in Eq. 4 is negligibly small when the variability of the mediator is low. To demonstrate this, we assume (without the loss of generality) that the ground-truth counterfactual mediator follows one of the BGM solutions, e.g.,* $\mathbb{P}(M_{a'} \mid X = x, A = a, M = m) = \delta(\mathbb{F}^{-1}(\mathbb{F}(m; x, a); x, a'))$*; and our GAN estimates another, i.e.,* $\mathbb{P}(M_{a'} \mid X = x, A = a, M = m) = \delta(\mathbb{F}^{-1}(1 - \mathbb{F}(m; x, a); x, a'))$*. Then, assuming a perfect fit of the GAN, the conditional expectation of the squared difference between the ground-truth counterfactual mediator and the estimated mediator is*

$$\sup_{G \in \mathcal{G}} \mathbb{E}\left[\left\|M_{A'} - \hat{M}_{A'}\right\|_2^2 \mid X = x, A = a\right] \tag{17}$$

$$= \mathbb{E}\left[\left|\mathbb{F}^{-1}(\mathbb{F}(M; x, a); x, a') - \mathbb{F}^{-1}(1 - \mathbb{F}(M; x, a); x, a')\right| \mid X = x, A = a\right] \tag{18}$$

$$= \mathbb{E}\left[\left|\mathbb{F}^{-1}(U; x, a') - \mathbb{F}^{-1}(1 - U; x, a')\right|\right] \tag{19}$$

$$= \int_0^1 \left|\mathbb{F}^{-1}(u; x, a') - \mathbb{F}^{-1}(1 - u; x, a')\right| \mathrm{d}u \tag{20}$$

$$\leq \int_0^1 \left|\mathbb{F}^{-1}(u; x, a') - \mu(x, a')\right| \mathrm{d}u + \int_0^1 \left|\mathbb{F}^{-1}(1 - u; x, a') - \mu(x, a')\right| \mathrm{d}u \tag{21}$$

$$= 2\,\mathbb{E}\left[\left|M - \mu(x, a')\right| \mid X = x, A = a'\right] \tag{22}$$

$$\overset{(*)}{\leq} 2\sqrt{\mathrm{Var}\left[M \mid X = x, A = a'\right]}, \tag{23}$$

where $(*)$ holds as an inequality between the mean absolute deviation and the standard deviation.

D.2. Results on the consistent estimation of the counterfactual distributions

Proposition 1 (Consistent estimation of the counterfactual distribution up to an MPI) *If the deterministic generator in Step 1 is a continuously differentiable function with respect to M, then it consistently estimates one of the counterfactual distributions of the mediator (Eq. 4).*

Proof

This can be proved in two steps. (i) We show that, given an optimal discriminator, the deterministic generator of our GAN estimates the distribution of potential mediators for counterfactual sensitive attributes, i.e., $\mathbb{P}(G(x, a, M_a)_{a'} \mid X = x, A = a) = \mathbb{P}(M_{a'} \mid X = x, A = a)$ in distribution. (ii) Then, we demonstrate that the outputs of the deterministic generator, conditional on the factual mediator $M = m$, estimate $\mathbb{P}(M_{a'} \mid X = x, A = a, M = m)$.

(i) Let $\pi_a(x) = \mathbb{P}(A = a \mid X = x)$ denote the propensity score. The discriminator of our GAN, given the covariates $X = x$, tries to distinguish between generated counterfactual data and ground truth factual data. The adversarial objective from Eq. 9 could be expanded with the law of total expectation wrt. X and A in the following way:

$$
\begin{aligned}
&\mathbb{E}_{(X,A,M)\sim\mathbb{P}_{\mathrm{f}}}\left[\log\left(D(X,\tilde{G}\left(X,A,M\right))_A\right)\right] && (24)\\
=&\mathbb{E}_{X\sim\mathbb{P}(X)}\mathbb{E}_{(A,M)\sim\mathbb{P}(A,M|X)}\left[\log\left(D(X,\tilde{G}\left(X,A,M\right))_A\right)\right] && (25)\\
=&\mathbb{E}_{X\sim\mathbb{P}(X)}\Big[\mathbb{E}_{M\sim\mathbb{P}(M|X,A=0)}\left[\log\left(D(X,\tilde{G}\left(X,0,M\right))_0\right)\right]\,\pi_0(X) && (26)\\
&\qquad\qquad+\mathbb{E}_{(M\sim\mathbb{P}(M|X,A=1)}\left[\log\left(D(X,\tilde{G}\left(X,1,M\right))_1\right)\right]\,\pi_1(X)\Big]\\
=&\mathbb{E}_{X\sim\mathbb{P}(X)}\Big[\mathbb{E}_{M\sim\mathbb{P}(M|X,A=0)}\left[\log\left(D(X,\{M,G\left(X,0,M\right)_1\})_0\right)\right]\,\pi_0(X) && (27)\\
&\qquad\qquad+\mathbb{E}_{M\sim\mathbb{P}(M|X,A=1)}\left[\log\left(1-D(X,\{G\left(X,1,M\right)_0,M\})_0\right)\right]\,\pi_1(X)\Big].
\end{aligned}
$$

We give a more detailed explanation of the derivation of this step. We denote $\{M, G\left(X, 0, M\right)_1\}$ in Eq. 26 as $\tilde{G}(X, 0, M)$. We modify the output of G before passing it as input to D. We replace the generated factual mediator $\hat{M}_A$ with the observed factual mediator M. We denote the new, combined data by $\tilde{G}\left(X, A, M\right)$. $\tilde{G}(X, 0, M)$ means the input of the discriminator for $A = 0$. In Eq. 7, we have defined $\tilde{G}(X, A, M)_a$ via

$$
\tilde{G}(X,A,M)_a = \begin{cases} M, & \text{if } A = a \\ G(X,A,M)_a, & \text{if } A = a' \end{cases} \tag{28}
$$

For the first term in Eq. 26, the mediator M is drawn from $\mathbb{P}(M \mid X, A = 0)$. For $A = 0$, we have

$$
\tilde{G}(X,0,M)_a = \begin{cases} M, & \text{if } A = a \\ G(X,0,M)_a, & \text{if } A = a' \end{cases} \tag{29}
$$

When $a = 0$, we have $\tilde{G}(X, 0, M)_0 = M$; when $a = 1$, we have $\tilde{G}(X, 0, M)_1 = G(X, 0, M)_1$. Therefore, we can write $\tilde{G}(X, 0, M) = \{M, G(X, 0, M)_1\}$. By replacing this term, we have shown that the first term in Eq. 26 equals the first term in Eq. 27.

Similarly, for $A = 1$, we have $\tilde{G}(X, 1, M) = \{G(X, 1, M)_0, M\}$, because, when $a = 0$ and $A = a'$, we have $\tilde{G}(X, 1, M)_0 = G(X, 1, M)_0$; when $a = 1$, we have $\tilde{G}(X, 0, M)_1 = M$.

The discriminator D then determines which component of $\tilde{G}$ is the observed factual mediator and thus outputs the corresponding probability, which is given by Eq. 8, i.e., $D(X, \tilde{G})_a = \hat{\mathbb{P}}(M = \tilde{G}_a \mid X, \tilde{G}) = \hat{\mathbb{P}}(A = a \mid X, \tilde{G})$. As it is the corresponding probability, therefore, the sum of the $D(X, \tilde{G})_0$ and $D(X, \tilde{G})_1$ should be 1. By replacing the term $\tilde{G}(X, 1, M) = \{G(X, 1, M)_0, M\}$ as we have shown above, we have

$$\log\left(D(X, \tilde{G}\left(X, 1, M\right))_1\right) = \log\left(1 - D(X, \{G\left(X, 1, M\right)_0, M\})_0\right) \tag{30}$$

Thus, we have shown that the second term in Eq. 26 equals the second term in Eq. 27.

Let $Z_0 = \{M, G\left(X, 0, M\right)_1\}$ and $Z_1 = \{G\left(X, 1, M\right)_0, M\}$ be two random variables. Then, using the law of the unconscious statistician, the expression can be converted to a weighted conditional GAN adversarial loss (Mirza and Osindero, 2014), i.e.,

$$\begin{aligned} &\mathbb{E}_{X\sim\mathbb{P}(X)}\Big[\mathbb{E}_{Z_0\sim\mathbb{P}(Z_0|X,A=0)}\left[\log\left(D(X, Z_0)_0\right)\right]\pi_0(X) \\ &\qquad + \mathbb{E}_{Z_1\sim\mathbb{P}(Z_1|X,A=1)}\left[\log\left(1 - D(X, Z_1)_0\right)\right]\pi_1(X)\Big] \end{aligned} \tag{31}$$

$$\begin{aligned} =&\mathbb{E}_{X\sim\mathbb{P}(X)}\Bigg[\int_{\mathcal{Z}}\Big(\log\left(D(X, z)_0\right)\pi_0(X)\mathbb{P}(Z_0 = z \mid X, A = 0) \\ &\qquad + \log\left(1 - D(X, z)_0\right)\pi_1(X)\mathbb{P}(Z_1 = z \mid X, A = 1)\Big)\,\mathrm{d}z\Bigg], \end{aligned} \tag{32}$$

where $\mathcal{Z} = \mathcal{M} \times \mathcal{M}$. Notably, the weights of the loss, i.e., $\pi_0(X)$ and $\pi_1(X)$, are greater than zero, due to the overlap assumption. The second term follows analogously. Following the theory from the standard GANs (Goodfellow, 2017), for any $(a, b) \in \mathbb{R}^2 \setminus 0$, the function $y \mapsto \log(y)a + \log(1-y)b$ achieves its maximum in $[0, 1]$ at $\frac{a}{a+b}$. Therefore, for a given generator, an optimal discriminator is

$$D(x, z)_0 = \frac{\mathbb{P}(Z_0 = z \mid X = x, A = 0)\,\pi_0(x)}{\mathbb{P}(Z_0 = z \mid X = x, A = 0)\,\pi_0(x) + \mathbb{P}(Z_1 = z \mid X = x, A = 1)\,\pi_1(x)}. \tag{33}$$

Both conditional densities used in the expression above can be expressed in terms of the potential outcomes densities due to the consistency and unconfoundedness assumptions, namely

$$\begin{aligned} \mathbb{P}(Z_0 = z \mid X = x, A = 0) &= \mathbb{P}(\{M = m_0, G(x, 0, M)_1 = m_1\} \mid X = x, A = 0) \\ &= \mathbb{P}(\{M_0 = m_0, G(x, 0, M_0)_1 = m_1\} \mid X = x), \end{aligned} \tag{34}$$

$$\begin{aligned} \mathbb{P}(Z_1 = z \mid X = x, A = 1) &= \mathbb{P}(\{G(x, 1, M)_0 = m_0, M = m_1\} \mid X = x, A = 1) \\ &= \mathbb{P}(\{G(x, 1, M_1)_0 = m_0, M_1 = m_1\} \mid X = x). \end{aligned} \tag{35}$$

Thus, an optimal generator of the GAN then minimizes the following conditional propensity-weighted Jensen–Shannon divergence (JSD)

$$\mathrm{JSD}_{\pi_0(x),\pi_1(x)}\Big(\mathbb{P}(\{M_0, G(x, 0, M_0)_1\} \mid X = x)\,\Big|\Big|\,\mathbb{P}(\{G(x, 1, M_1)_0, M_1\} \mid X = x)\Big), \tag{36}$$

where $\mathrm{JSD}_{w_1,w_2}(\mathbb{P}_1 \,||\, \mathbb{P}_1) = w_1\,\mathrm{KL}(\mathbb{P}_1 \,||\, w_1\,\mathbb{P}_1 + w_2\,\mathbb{P}_2) + w_2\,\mathrm{KL}(\mathbb{P}_2 \,||\, w_1\,\mathbb{P}_1 + w_2\,\mathbb{P}_2)$ and where $\mathrm{KL}(\mathbb{P}_1 \,||\, \mathbb{P}_1)$ is Kullback–Leibler divergence. The Jensen–Shannon divergence is minimized when $G(x, 0, M_0)_1 = M_1$ and $G(x, 1, M_1)_0 = M_0$ conditioned on $X = x$ (in distribution), since, in this case, it equals zero, i.e.,

$$\mathbb{P}(G(x, a, M_a)_{a'} \mid X = x) = \mathbb{P}(M_{a'} \mid X = x). \tag{37}$$

Finally, due to the unconfoundedness assumption, the generator of our GAN estimates the potential mediator distributions with counterfactual sensitive attributes, i.e.,

$$\mathbb{P}(G(x, a, M_a)_{a'} \mid X = x, A = a) = \mathbb{P}(M_{a'} \mid X = x, A = a) \tag{38}$$

in distribution.

(ii) For a given factual observation, $X = x, A = a, M = m$, our generator yields a deterministic output, i.e.,

$$\begin{aligned}\mathbb{P}(G(x, a, M_a)_{a'} \mid X = x, A = a, M = m) &= \mathbb{P}(G(x, a, m)_{a'} \mid X = x, A = a, M = m) &(39)\\ &= \delta(G(x, a, m)_{a'}). &(40)\end{aligned}$$

At the same time, this counterfactual distribution is connected with the potential mediators' distributions with counterfactual sensitive attributes, $\mathbb{P}(M_{a'} = m' \mid X = x, A = a)$, via the law of total probability:

$$\begin{aligned}&\mathbb{P}(M_{a'} = m' \mid X = x, A = a) = \mathbb{P}(G(x, a, M)_{a'} = m' | X = x, A = a) &(41)\\ &= \int_{\mathcal{M}} \delta(G(x, a, m)_{a'} - m')\,\mathbb{P}(M = m \mid X = x, A = a)\,\mathrm{d}m &(42)\\ &= \sum_{m:\, G(x,a,m)_{a'} = m'} |\,\nabla_m G(x, a, m)_{a'}\,|^{-1}\,\mathbb{P}(M = m \mid X = x, A = a). &(43)\end{aligned}$$

Due to the unconfoundedness and the consistency assumptions, this is equivalent to

$$\mathbb{P}(M = m' \mid X = x, A = a') = \sum_{m:\, G(x,a,m)_{a'} = m'} |\,\nabla_m G(x, a, m)_{a'}\,|^{-1}\,\mathbb{P}(M = m \mid X = x, A = a). \tag{44}$$

The equation above has only two solutions wrt. $G(x, a, \cdot)$ in the class of the continuously differentiable functions (Corollary 3 in Melnychuk et al. (2023)), namely:[9]

$$G(x, a, m)_{a'} = \mathbb{F}^{-1}(\pm\mathbb{F}(m; x, a) \mp 0.5 + 0.5; x, a'), \tag{45}$$

where $\mathbb{F}(\cdot; x, a)$ and $\mathbb{F}^{-1}(\cdot; x, a)$ are a CDF and an inverse CDF of $\mathbb{P}(M \mid X = x, A = a)$. Thus, the deterministic generator of GAN exactly matches one of the two BGM solutions from (i). This concludes that our deterministic generator consistently estimates the counterfactual distribution of the mediator, $\mathbb{P}(M_{a'} \mid X = x, A = a, M = m)$, up to a measure-preserving indeterminacy. ■

9. Under mild conditions, the counterfactual distributions cannot be defined via the point mass distribution with non-monotonous functions, even if we assume the extension of BGMs to all non-monotonous continuously differentiable functions.

Corollary 1 (MPI identifiability of high-dimensional mediators) *The results of the Proposition 1 generalize to multi-dimensional mediators. In this case, the deterministic generator consistently estimates one of the implicitly given counterfactual distributions:*

$$\mathbb{P}(M_{a'} = m' \mid X = x, A = a) = \int_{\mathcal{M}} \delta(G(x, a, m)_{a'} - m')\, \mathbb{P}(M = m \mid X = x, A = a)\, \mathrm{d}m. \tag{46}$$

Corollary 4 *The results of the Proposition 1 naturally generalize to sensitive attributes with more categories, i.e., $\mathcal{A} = \{0, 1, \ldots, k-1\}, k > 2$.*

Proof We want to show that, when $\mathcal{A} = \{0, 1, \ldots, k-1\}, k > 2$, the deterministic generator is still able to learn the potential mediator distributions with the counterfactual distributions. For that, we follow the same derivation steps as in part (i) of the proof of Proposition 1. This brings us to the following equality for the loss of the discriminator:

$$\begin{aligned}
&\mathbb{E}_{(X,A,M)\sim\mathbb{P}_{\mathrm{f}}}\left[\log\left(D(X, \tilde{G}\,(X, A, M))_A\right)\right] && (47)\\
=&\mathbb{E}_{X\sim\mathbb{P}(X)}\Bigg[\int_{\mathcal{Z}}\Big(\log\left(D(X,z)_0\right)\pi_0(X)\mathbb{P}(Z_0 = z \mid X, A = 0) && (48)\\
&\qquad + \log\left(D(X,z)_1\right)\pi_1(X)\mathbb{P}(Z_1 = z \mid X, A = 1) && (49)\\
&\qquad \ldots && (50)\\
&\qquad + \log\left(D(X,z)_{k-2}\right)\pi_{k-2}(X)\mathbb{P}(Z_{k-2} = z \mid X, A = k-2) && (51)\\
&\qquad + \log\Big(1 - \sum_{j=0}^{k-2} D(X,z)_j\Big)\pi_{k-1}(X)\mathbb{P}(Z_{k-1} = z \mid X, A = k-1)\Big)\,\mathrm{d}z\Bigg], && (52)
\end{aligned}$$

where

$$\begin{aligned}
Z_0 &= \{M, G(X,0,M)_1, G(X,0,M)_2, \ldots, G(X,0,M)_{k-1})\}, && (53)\\
Z_1 &= \{G(X,1,M)_0, M, G(X,1,M)_2, \ldots, G(X,1,M)_{k-1}\}, && (54)\\
&\ldots && (55)\\
Z_{k-1} &= \{G(X,k-1,M)_0, G(X,k-1,M)_1, \ldots, M\}. && (56)
\end{aligned}$$

Then, it is easy to see that, for a given generator, an optimal discriminator is (analogously to Eq. 33)

$$D(x,z)_a = \frac{\mathbb{P}(Z_a = z \mid X = x, A = a)\,\pi_a(x)}{\sum_{j=0}^{k-1}\mathbb{P}(Z_j = z \mid X = x, A = j)\,\pi_j(x)} \quad \text{for all } a \in \mathcal{A}. \tag{57}$$

This happens, as, for any $(a_0, \ldots, a_{k-1}) \in \mathbb{R}^k \setminus 0$, the function $(y_0, y_1, \ldots y_{k-2}) \mapsto \log(y_0)a_0 + \log(y_1)a_1 + \cdots + \log(y_{k-2})a_{k-2} + \log(1 - \sum_{j=0}^{k-2} y_j)a_{k-1}$ achieves its maximum in $[0,1]$ at $\left(\frac{a_0}{\sum_{j=0}^{k-1} a_j}, \frac{a_1}{\sum_{j=0}^{k-1} a_j}, \ldots, \frac{a_{k-2}}{\sum_{j=0}^{k-1} a_j}\right)$. Then, an optimal generator of the GAN aims to minimize the propensity-weighted multi-distribution JSD, i.e.,

$$\begin{aligned}
\mathrm{JSD}_{\pi_0(x),\pi_1(x),\ldots,\pi_{k-1}(x)}\Big(&\mathbb{P}(\{M_0, G(x,0,M_0)_1, G(x,0,M_0)_2, \ldots, G(x,0,M_0)_{k-1}\} \mid X = x),\\
&\mathbb{P}(\{G(x,1,M_1)_0, M_1, G(x,1,M_1)_2, \ldots, G(x,1,M_1)_{k-1}\} \mid X = x),\\
&\ldots\\
&\mathbb{P}(\{G(x,k-1,M_{k-1})_0, G(x,k-1,M_{k-1})_1, \ldots, M_{k-1}\} \mid X = x)\Big).
\end{aligned} \tag{58}$$

The JSD is minimized when all the distributions are equal. If we additionally look at the marginalized distributions, the following equalities will hold

$$\mathbb{P}(G(x, a, M_a)_{a'} \mid X = x) = \mathbb{P}(M_{a'} \mid X = x) \quad \text{for all } a \neq a' \in \mathcal{A}. \tag{59}$$

This concludes the proof of the Corollary, as all additional steps are analogous to the Proposition 1. ■

D.3. Proof of Proposition 2

Here, we prove Proposition 2 from the main paper, which states that our counterfactual regularization achieves counterfactual fairness if our deterministic generator consistently estimates the counterfactuals.

Proposition 2 (Counterfactual mediator regularization bound) *Given the prediction model h that is Lipschitz continuous with a Lipschitz constant $\mathcal{C}$, we have*

$$\mathbb{E}\left[\|(h(X, M) - h(X, M_{A'})\|_2^2\right] \leq \mathcal{C}\,\mathbb{E}\left[\left\|M_{A'} - \hat{M}_{A'}\right\|_2^2\right] + \sup_{G \in \mathcal{G}} \mathcal{R}_{\text{cm}}(h, G), \text{ for every } G \in \mathcal{G}, \tag{60}$$

where $\hat{M}_{A'} = G(X, A, M)_{A'}$ and $\mathcal{G}$ is a set of all the deterministic generators, minimizing the Eq. 10.

Proof Using the triangle inequality, we yield

$$\begin{aligned}
&\mathbb{E}\left[\|h(X, M) - h(X, M_{A'})\|_2^2\right] && (61)\\
=&\mathbb{E}\left[\left\|h(X, M) - h(X, M_{A'}) + h(X, \hat{M}_{A'}) - h(X, \hat{M}_{A'})\right\|_2^2\right] && (62)\\
\leq&\mathbb{E}\left[\left\|h(X, M) - h(X, \hat{M}_{A'})\right\|_2^2\right] + \mathbb{E}\left[\left\|h(X, \hat{M}_{A'}) - h(X, M_{A'})\right\|_2^2\right] && (63)\\
=&\mathbb{E}\left[\left\|h(X, \hat{M}_{A'}) - h(X, M_{A'})\right\|_2^2\right] + \mathcal{R}_{\text{cm}}(h, G) && (64)\\
\leq&\mathcal{C}\,\mathbb{E}\left[\left\|(X, \hat{M}_{A'}) - (X, M_{A'})\right\|_2^2\right] + \mathcal{R}_{\text{cm}}(h, G) && (65)\\
=&\mathcal{C}\,\mathbb{E}\left[\left\|M_{A'} - \hat{M}_{A'}\right\|_2^2\right] + \mathcal{R}_{\text{cm}}(h, G) && (66)\\
\leq&\mathcal{C}\,\mathbb{E}\left[\left\|M_{A'} - \hat{M}_{A'}\right\|_2^2\right] + \sup_{G \in \mathcal{G}} \mathcal{R}_{\text{cm}}(h, G). && (67)
\end{aligned}$$

■

Appendix E. Training algorithm of GCFN

Algorithm 1 Training algorithm of GCFN

Input: Training dataset $\mathcal{D}$; fairness weight λ; number of training GANs to train s; number of training epoch for each GAN e_1; number of training prediction model epoch e_2; minibatch of size n; training supervised loss weight α.
Init: Generator G parameters: θ_g; discriminator D parameters: θ_d; prediction model h parameters: θ_h.
Output: Counterfactually fair prediction model h.

Step 1: Training GAN to learn to generate counterfactual mediator
for $j \leftarrow 1$ **to** s **do**
  for *epoch* $\leftarrow 1$ **to** e_1 **do**
    for $t \leftarrow 1$ **to** k **do** `// Train discriminator` D_j
      Sample minibatch $\{x^{(i)}, a^{(i)}, m^{(i)}\}_{i=1}^n$ from $\mathcal{D}$
      Compute generator output: $G_j\left(x^{(i)}, a^{(i)}, m^{(i)}\right)_a = \hat{m}_a^{(i)} \quad \text{for } a \in \{0,1\}$
      Modify G_j output to: $\tilde{G}_j^{(i)}(a) = \begin{cases} m^{(i)}, & \text{if } a^{(i)} = a, \\ \hat{m}_a^{(i)}, & \text{if } a^{(i)} = a' \end{cases} \quad \text{for } a \in \{0,1\}$
      Update discriminator via stochastic gradient ascent: $\nabla_{\theta_d} \frac{1}{n} \sum_{i=1}^n \log\left(D_j(x^{(i)}, \tilde{G}_j^{(i)})_{a^{(i)}}\right)$
    end
  end
  for $t \leftarrow 1$ **to** k **do** `// Train generator` G_j
    Sample minibatch $\{x^{(i)}, a^{(i)}, m^{(i)}\}_{i=1}^n$ from $\mathcal{D}$
    Compute generator output: $G_j\left(x^{(i)}, a^{(i)}, m^{(i)}\right)_a = \hat{m}_a^{(i)} \quad \text{for } a \in \{0,1\}$
    Modify G_j output to: $\tilde{G}_j^{(i)}(a) = \begin{cases} m^{(i)}, & \text{if } a^{(i)} = a, \\ \hat{m}_a^{(i)}, & \text{if } a^{(i)} = a' \end{cases} \quad \text{for } a \in \{0,1\}$
    Update generator via stochastic gradient descent:

$$\nabla_{\theta_g} \frac{1}{n} \sum_{i=1}^n \left[\log\left(D_j(x^{(i)}, \tilde{G}_j^{(i)})_{a^{(i)}}\right) + \log\left(1 - D_j(x^{(i)}, \tilde{G}_j^{(i)})_{1-a^{(i)}}\right) + \alpha \left\| m^{(i)} - G_j(x^{(i)}, a^{(i)}, m^{(i)})_{a^{(i)}} \right\|_2^2 \right]$$

  end
end

Step 2: Training prediction model with counterfactual mediator regularization
for *epoch* $\leftarrow 1$ **to** e_2 **do**
  Sample minibatch $\{x^{(i)}, a^{(i)}, m^{(i)}, y^{(i)}\}_{i=1}^n$ from $\mathcal{D}$
  Generate $\hat{m}_j^{(i)}$ from $G_j(x^{(i)}, a^{(i)}, m^{(i)})$
  Compute counterfactual mediator regularization: $\mathcal{R}_{\text{cm}}(h, G_j) = \left\| h(x^{(i)}, m^{(i)}) - h(x^{(i)}, \hat{m}_{j,a'^{(i)}}^{(i)}) \right\|_2^2$
  Update prediction model via stochastic gradient descent:

$$\nabla_{\theta_h} \frac{1}{n} \sum_{i=1}^n \left[y^{(i)} \log(h(x^{(i)}, m^{(i)})) + (1 - y^{(i)}) \log(1 - h(x^{(i)}, m^{(i)})) + \lambda \max_{j=1}^s \mathcal{R}_{\text{cm}}(h, G_j) \right]$$

end

Appendix F. Generalization to multiple social groups

F.1. Step 1: Generating counterfactuals of the mediator

Our method can easily be extended to scenarios with multiple social groups. Suppose we have k categories, then the sensitive attribute $A \in \mathcal{A}$, where $\mathcal{A} = \{0, 1, \ldots, k-1\}$ and $k > 2$.

The output of the deterministic generator G is k potential mediators, i.e., $\hat{M}_0, \hat{M}_1, \ldots, \hat{M}_{k-1}$, from which one is factual, and the others are counterfactual, i.e.,

$$G(X, A, M)_a = \hat{M}_a \quad \text{for} \quad a \in \{0, 1, \ldots, k-1\} \tag{68}$$

The reconstruction loss of the deterministic generator is the same as the binary case:

$$\mathcal{L}_{\mathrm{f}}(G) = \mathbb{E}_{(X,A,M)\sim\mathbb{P}_{\mathrm{f}}}\left[\|M - G(X, A, M)_A\|_2^2\right], \tag{69}$$

where $\|\cdot\|_2$ is the L_2-norm.

The discriminator D is designed to differentiate the factual mediator M (as observed in the data) from the $k-1$ generated counterfactual mediators (as generated by G).

We modify the output of G before passing it as input to D: We replace the generated factual mediator $\hat{M}_A$ with the observed factual mediator M. We denote the new, combined data by $\tilde{G}(X, A, M)$, which is defined via

$$\tilde{G}(X, A, M)_a = \begin{cases} M, & \text{if } A = a, \\ G(X, A, M)_a, & \text{Otherwise}, \end{cases} \quad \text{for } a \in \{0, 1, \ldots, k-1\}. \tag{70}$$

The discriminator D then determines which component of $\tilde{G}$ is the observed factual mediator and thus outputs the corresponding probability. Formally, for the input $(X, \tilde{G})$, the output of the discriminator D is

$$D\left(X, \tilde{G}\right)_a = \hat{\mathbb{P}}\left(M = \tilde{G}_a \mid X, \tilde{G}\right) = \hat{\mathbb{P}}\left(A = a \mid X, \tilde{G}\right) \quad \text{for } a \in \{0, 1, \ldots, k-1\}. \tag{71}$$

Step 1 is trained in an adversarial manner: (i) the deterministic generator G seeks to generate counterfactual mediators in a way that minimizes the probability that the discriminator can differentiate between factual mediators and counterfactual mediators, while (ii) the discriminator D seeks to maximize the probability of correctly identifying the factual mediator. We thus use an adversarial loss $\mathcal{L}_{\mathrm{adv}}$ by

$$\mathcal{L}_{\mathrm{adv}}(G, D) = \mathbb{E}_{(X,A,M)\sim\mathbb{P}_{\mathrm{f}}}\left[\log\left(D(X, \tilde{G}(X, A, M))_A\right)\right]. \tag{72}$$

Overall, our Step 1 is trained through an adversarial training procedure with a minimax problem as

$$\min_G \max_D \mathcal{L}_{\mathrm{adv}}(G, D) + \alpha \mathcal{L}_{\mathrm{f}}(G), \tag{73}$$

with a hyperparameter α on $\mathcal{L}_{\mathrm{f}}$.

F.2. Step 2: Counterfactual fair prediction through counterfactual mediator regularization

We use the output of Step 1 to train a prediction model h under counterfactual fairness in a supervised way. Our counterfactual mediator regularization $\mathcal{R}_{\mathrm{cm}}(h, G)$ thus is

$$\mathcal{R}_{\mathrm{cm}}(h, G) = \mathbb{E}_{(X,A,M)\sim\mathbb{P}_{\mathrm{f}}}\left[\frac{1}{(k-1)}\sum_{\substack{a=0\\a\neq A}}^{k-1}\left\|h\left(X, M\right) - h\left(X, \hat{M}_a\right)\right\|_2^2\right]. \tag{74}$$

The training loss is

$$\mathcal{L}(h) = \mathcal{L}_{\mathrm{ce}}(h) + \lambda \sup_{G\in\mathcal{G}} \mathcal{R}_{\mathrm{cm}}(h, G). \tag{75}$$

F.3. Theoretical insights

We provide proof that our method can naturally generalize to sensitive attributes with more categories in Appendix D, Corollary. 4.

Appendix G. Datasets

G.1. Synthetic data

Analogous to prior works that simulate synthetic data for benchmarking (Kim et al., 2021; Kusner et al., 2017; Quinzan et al., 2024), we generate our synthetic dataset in the following way. The covariates X is drawn from a standard normal distribution $\mathcal{N}(0,1)$. The sensitive attribute A follows a Bernoulli distribution with probability p, determined by a sigmoid function σ of X and a Gaussian noise term U_A. We then generate the mediator M as a function of X, A, and a Gaussian noise term U_M. Finally, the target Y follows a Bernoulli distribution with probability p_y, calculated by a sigmoid function of X, M, and a Gaussian noise term U_Y. β_i ($i \in [1,6]$) are the coefficients. Let $\sigma(x) = \frac{1}{1+\mathrm{e}^{-x}}$ represent the sigmoid function. Formally, we yield

$$\begin{cases} X = U_X, & U_X \sim \mathcal{N}(0,1), \\ A \sim \text{Bernoulli}\left(\sigma(\beta_1 X + U_A)\right), & U_A \sim \mathcal{N}(0,0.01), \\ M = \beta_2 X + \beta_3 A + U_M, & U_M \sim \mathcal{N}(0,0.01), \\ Y \sim \text{Bernoulli}\left(\sigma(\beta_5 X + \beta_6 M + U_Y)\right), & U_Y \sim \mathcal{N}(0,0.01). \end{cases} \tag{76}$$

We sample 10,000 observations and use 20% as the test set.

G.2. Semi-synthetic data

LSAC dataset. The Law School (LSAC) dataset (Wightman, 1998) contains information about the law school admission records. We use the LSAC dataset to construct two semi-synthetic datasets. In both, we set the sensitive attribute to *gender*. We take *resident* and *race* from the LSAC dataset as confounding variables. The *LSAT* and *GPA* are the mediator variables, and the *admissions decision* is our target variable. We simulate 101,570 samples and use 20% as the test set. We denote M_1 as GPA score, M_2 as LSAT score, X_1 as resident, and X_2 as race. Further, $w_{X_1}, w_{X_2}, w_A, w_{M_1}, w_{M_2}$ are the coefficients. U_{M_1}, U_{M_2}, U_Y are the Gaussian noise. Let $\sigma(x) = \frac{1}{1+\mathrm{e}^{-x}}$ represent the sigmoid function.

We follow the prior work (Bica et al., 2020) to produce two different semi-synthetic datasets as follows. For the first one, we use the sigmoid function on linear combinations, and, for the second one, we use the sinus function that could make extrapolation more challenging for our GCFN.

■ Semi-synthetic dataset "sigmoid":

$$\begin{cases} M_1 = w_{M_1}\left(\sigma(w_A A + w_{X_1} X_1 + w_{X_2} X_2 + U_{M_1})\right), & U_{M_1} \sim \mathcal{N}(0,0.01), \\ M_2 = w_{M_2} + w_{M_1}\left(\sigma(w_A S + w_{X_1} X_1 + w_{X_2} X_2 + U_{M_2})\right), & U_{M_2} \sim \mathcal{N}(0,0.01), \\ Y \sim \text{Bernoulli}\left(\sigma(w_{M_1} M_1 + w_{M_2} M_2 + w_{X_1} X_1 + w_{X_2} X_2 + U_Y)\right), & U_Y \sim \mathcal{N}(0,0.01). \end{cases} \tag{77}$$

■ Semi-synthetic "sin":

$$\begin{cases} M_1 = w_A \cdot A - \sin\left(\pi \times (w_{X_1} X_1 + w_{X_2} X_2 + U_{M_1})\right), & U_{M_1} \sim \mathcal{N}(0,0.01), \\ M_2 = w_A \cdot A - \sin\left(\pi \times (w_{X_1} X_1 + w_{X_2} X_2 + U_{M_2})\right), & U_{M_2} \sim \mathcal{N}(0,0.01), \\ Y \sim \text{Bernoulli}\left(\sigma(w_{M_1} M_1 + w_{M_2} M_2 + w_{X_1} X_1 + w_{X_2} X_2 + U_Y)\right), & U_Y \sim \mathcal{N}(0,0.01). \end{cases} \tag{78}$$

G.3. Real-world data

UCI Adult dataset: The UCI Adult dataset (Asuncion and Newman, 2007) captures information about 48,842 individuals including their sociodemographics. Our aim is to predict if individuals earn more than USD 50k per year. We follow the setting of earlier research (Kim et al., 2021; Nabi and Shpitser, 2018; Quinzan et al., 2024; Xu et al., 2019). We treat *gender* as the sensitive attribute and set mediator variables to be *marital status*, *education level*, *occupation*, *hours per week*, and *work class*. The causal graph of the UCI dataset is in Fig. 4. We take 20% as the test set.

COMPAS dataset: COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) (Angwin et al., 2016) was developed as a decision support tool to score the likelihood of a person's recidivism. The score ranges from 1 (lowest risk) to 10 (highest risk). The dataset further contains information about whether there was an actual recidivism (reoffended) record within 2 years after the decision. Overall, the dataset has information about over 10,000 criminal defendants in Broward County, Florida. We treat *race* as the sensitive attribute. The mediator variables are the features related to prior convictions and the current charge degree. The target variable is the *recidivism* for each defendant. The causal graph of the COMPAS dataset is in Fig. 5. We take 20% as test set.

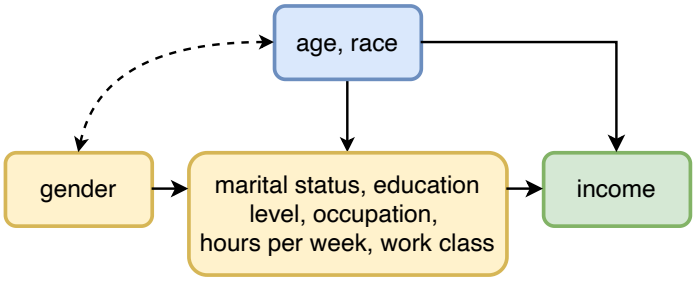

Figure 4: Causal graph of UCI dataset.

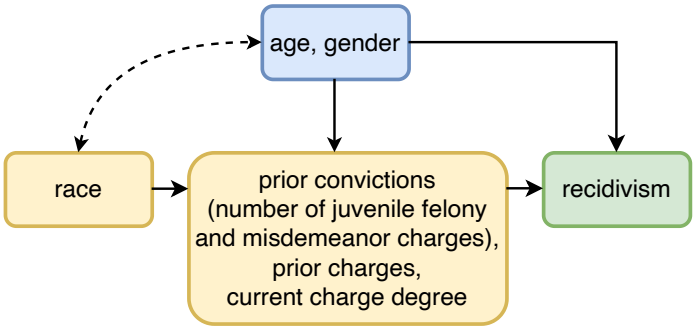

Figure 5: Causal graph of COMPAS dataset.

In practice, it is common and typically straightforward to choose which variables act as mediators M through domain knowledge (Nabi and Shpitser, 2018; Kim et al., 2021; Plečko and Bareinboim, 2024). Hence, mediators M are simply all variables that can potentially be influenced by the sensitive

attribute. All other variables (except for A and Y) are modeled as covariates X. For example, consider a job application setting where we want to avoid discrimination by gender. Then A is gender, and Y is the job offer. Mediators are, for instance, education level or work experience, as both are potentially influenced by gender. In contrast, age is a covariate because it is not influenced by gender.

Appendix H. Implementation Details

H.1. Baselines

We compare our method against the following state-of-the-art approaches: (1) **CFAN** (Kusner et al., 2017): the conceptual algorithm with additive noise where only non-descents of sensitive attributes and the estimated latent variables are used for prediction; (2) **CFUA** (Kusner et al., 2017): a variant of the algorithm which does not use the sensitive attribute or any descents of the sensitive attribute; (3) **mCEVAE** (Pfohl et al., 2019): adds a maximum mean discrepancy to regularize the generations in order to remove the information the inferred latent variable from sensitive information; (4) **DCEVAE** (Kim et al., 2021): a VAE-based approach that disentangles the exogenous uncertainty into two variables; (5) **ADVAE** (Grari et al., 2023): adversarial neural learning approach which should be more powerful than penalties from maximum mean discrepancy but is aimed the continuous setting; (6) **HSCIC** (Quinzan et al., 2024): originally designed to enforces the predictions to remain invariant to changes of sensitive attributes using conditional kernel mean embeddings but which we adapted for counterfactual fairness. We also adapt applicable baselines from fair dataset generation: (7) **CFGAN** (Xu et al., 2019): which we extend with a second-stage prediction model. Details are in Appendix H.

H.2. Implementation of our method

To ensure our GCFN to achieve counterfactual fairness, we train $s = 10$ GANs and consider the worst-case counterfactual fairness, i.e., we take the maximum value of counterfactual mediator regularization $\max_{j=1}^{s} \mathcal{R}_{\text{cm}}(h, G_j)$. Our GCFN is implemented in PyTorch. Both the deterministic generator and the discriminator in the GAN model are designed as deep neural networks, each with a hidden layer of dimension 64. LeakyReLU is employed as the activation function, and batch normalization is applied in the deterministic generator to enhance training stability. The GAN training procedure is performed for 300 epochs with a batch size of 256 at each iteration. We set the learning rate to 0.0005. Following the GAN training, the prediction model, structured as a multilayer perceptron (MLP), is trained separately. This classifier can incorporate spectral normalization in its linear layers to ensure Lipschitz continuity. It is trained for 30 epochs, with the same learning rate of 0.005 applied. The training time of our GCFN on (semi-) synthetic datasets is comparable to or smaller than the baselines.

H.3. Implementation of benchmarks

We implement **CFAN** (Kusner et al., 2017) in PyTorch based on the paper's source code in R and Stan on https://github.com/mkusner/counterfactual-fairness. We use a VAE to infer the latent variables. For **mCEVAE** (Pfohl et al., 2019), we follow the implementation from https://github.com/HyemiK1m/DCEVAE/tree/master/Tabular/mCEVAE_baseline. We implement **CFGAN** (Xu et al., 2019) in PyTorch based on the code of Abroshan et al. (2022) and the TensorFlow source code of Xu et al. (2019). We implement **ADVAE** (Grari et al., 2023) in PyTorch. For **DCEVAE** (Kim et al., 2021), we use the source code of the author of **DCEVAE** (Kim et al., 2021). We use **HSCIC** (Quinzan et al., 2024) source implementation from the supplementary material provided on the OpenReview website https://openreview.net/forum?id=ERjQnrmLKH4. We performed rigorous hyperparameter tuning for all baselines.

H.4. Hyperparameter tuning.

We perform a rigorous procedure to optimize the hyperparameters for the different methods as follows. For DCEVAE (Kim et al., 2021) and mCEVAE (Pfohl et al., 2019), we follow the hyperparameter optimization as described in the supplement of Kim et al. (2021). For ADVAE (Grari et al., 2023) and CFGAN (Xu et al., 2019), we follow the hyperparameter optimization as described in their paper. For both HSCIC and our GCFN, we have an additional weight that introduces a trade-off between accuracy and fairness. This provides additional flexibility to decision-makers as they tailor the methods based on the fairness needs in practice (Quinzan et al., 2024). We then benchmark the utility of different methods across different choices of γ of the utility function in Sec. 5. This allows us thus to optimize the trade-off weight λ inside HSCIC and our GCFN using grid search. For HSCIC, we experiment with $\lambda = 0.1, 0.5, 1, 5, 10, 15, 20$ and choose the best for them across different datasets. For our method, we experiment with $\lambda = 0.1, 0.5, 1, 1.5, 2$. Since the utility function considers two metrics, across the experiments on (semi-)synthetic datasets, the weight λ is set to 0.5 to get a good balance for our method.

Appendix I. Additional experimental results

I.1. Results for synthetic dataset

Setting: We follow previous works that simulate a fully synthetic dataset for performance evaluations (Kim et al., 2021; Quinzan et al., 2024). The details of the data generation process are in Appendix G. We generate 10,000 samples and use 20% as the test set.

Results: Results are shown in Fig. 6. We again find that our method is highly effective.

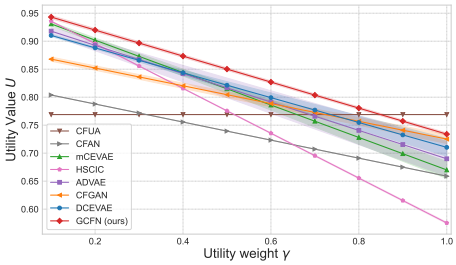

Figure 6: Results for synthetic datasets. A larger utility is better. Shown: mean $\pm$ std over 5 runs.

I.2. Results for (semi-)synthetic dataset

We compute the average value of the utility function U over varying utility weights $\gamma \in \{0.1, \ldots, 1.0\}$ on the synthetic dataset (Fig. 7) and two different semi-synthetic datasets (Fig. 8 and Fig. 9).

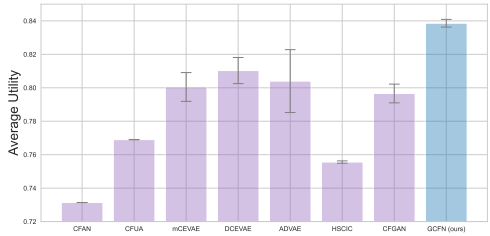

Figure 7: Average utility function value U across different utility weights γ on synthetic dataset.

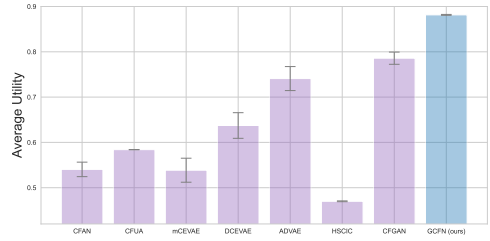

Figure 8: Average utility function value U across different utility weight γ on semi-synthetic (sigmoid) dataset.

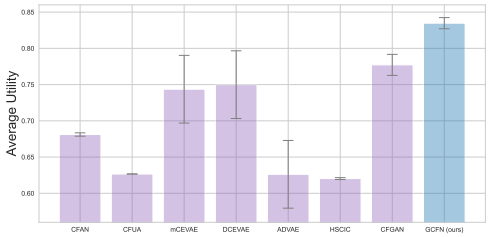

Figure 9: Average utility function value U across different utility weight γ on semi-synthetic (sin) dataset.

I.3. Results for UCI Adult dataset

We now examine the results for different fairness weights λ. For this, we report results from $\lambda = 0.5$ (Fig. 10) to $\lambda = 1000$ (Fig. 13). In line with our expectations, we see that larger values for fairness weight λ lead the distributions of the predicted target to overlap more, implying that counterfactual fairness is enforced more strictly. This shows that our regularization $\mathcal{R}_{\text{cm}}$ achieves the desired behavior.

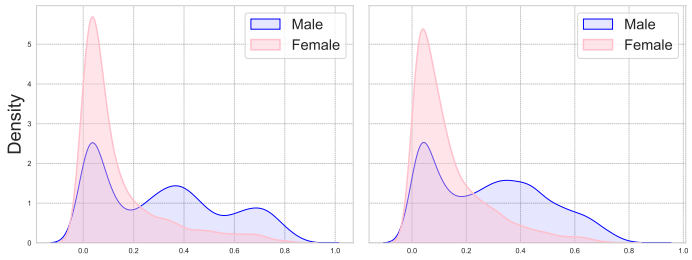

Figure 10: Density plots of the predicted target on the UCI Adult dataset. Left: fairness weight $\lambda = 0$. Right: fairness weight $\lambda = 0.5$.

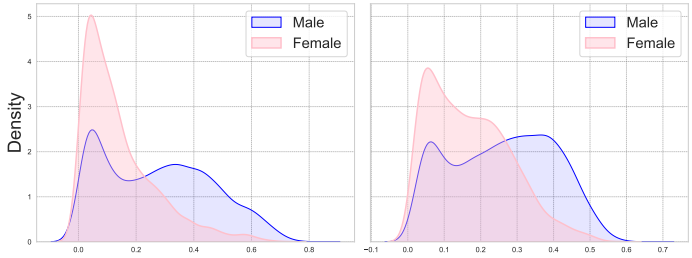

Figure 11: Density plots of the predicted target on the UCI Adult dataset. Left: fairness weight $\lambda = 1$. Right: fairness weight $\lambda = 5$.

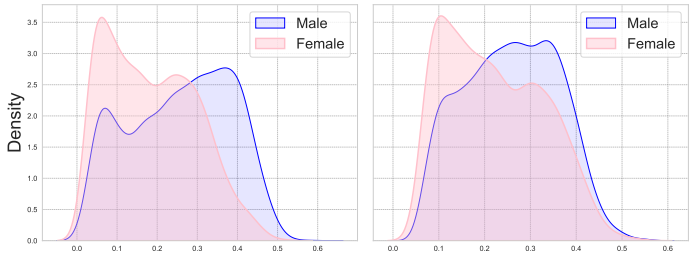

Figure 12: Density plots of the predicted target on the UCI Adult dataset. Left: fairness weight $\lambda = 10$. Right: fairness weight $\lambda = 100$.

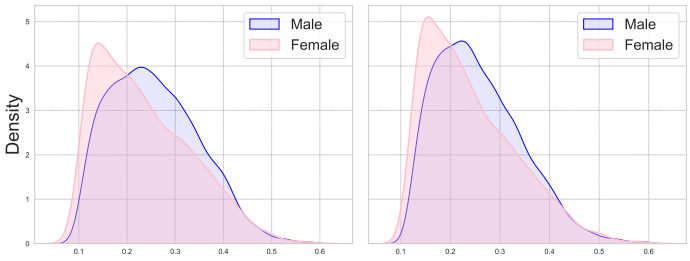

Figure 13: Density plots of the predicted target on the UCI Adult dataset. Left: fairness weight $\lambda = 500$. Right: fairness weight $\lambda = 1000$.

I.4. Results for COMPAS dataset

In Sec. 5, we show how black defendants are treated differently by the COMPAS score vs. our GCFN. Here, we also show how white defendants are treated differently by the COMPAS score vs. our GCFN; see Fig. 14. We make the following observations. (1) Our GCFN makes oftentimes different predictions for white defendants with a low and high COMPAS score, which is different from black defendants. (2) Our method also arrives at different predictions for white defendants with low prior charges, similar to black defendants.

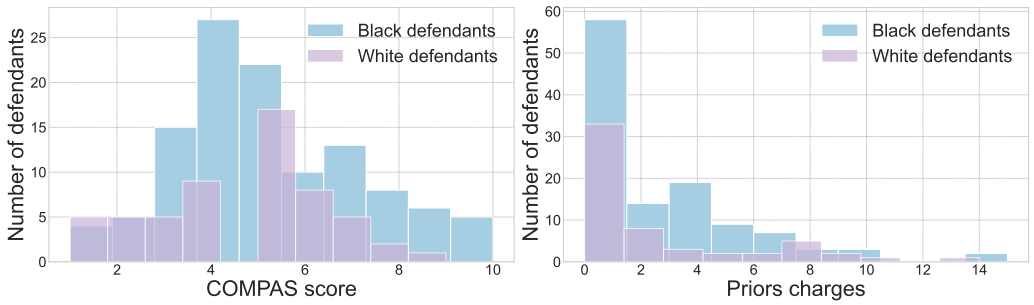

Figure 14: Distribution of white and black defendants that are treated differently using our GCFN. Left: COMPAS score. Right: Prior charges.

Appendix J. Additional insights: Why our method does not copy factuals? Why does it learn counterfactuals?

As an additional analysis, we now provide further insights into how our GCFN operates. Specifically, one may think that our GCFN simply learns to reproduce factual mediators in the GAN rather than actually learning the counterfactual mediators. However, this is *not* the case. To show this, we compare the (1) the factual mediator M, (2) the ground-truth counterfactual mediator $M_{A'}$, and (3) the generated counterfactual mediator $\hat{M}_{A'}$. The normalized mean squared error (MSE) between them is in Table 2. We find: (1) The factual mediator and the generated counterfactual mediator are highly dissimilar. This is shown by a normalized MSE($M, \hat{M}_{A'}$) of ≈ 1. (2) The ground-truth counterfactual mediator and our generated counterfactual mediator are highly similar. This shown by a normalized MSE($M_{A'}, \hat{M}_{A'}$) of close to zero. In sum, our GCFN is effective in learning counterfactual mediators (and does *not* reproduce the factual data).

Table 2: Our GCFN can learn the distribution of the counterfactual mediator. The normalized MSE($M_{A'}$,$\hat{M}_{A'}$) is ≈ 0, showing the generated counterfactual mediator is similar to the ground-truth counterfactual mediator. In contrast, both the factual and the generated counterfactual mediator are highly dissimilar.

	Synthetic	Semi-syn. (sigmoid)	Semi-syn. (sin)
MSE(M, $M_{A'}$)	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$
MSE(M, $\hat{M}_{A'}$)	$1.21_{\pm 0.064}$	$1.02_{\pm 0.027}$	$1.06_{\pm 0.051}$
MSE($M_{A'}$, $\hat{M}_{A'}$)	$0.14_{\pm 0.052}$	$0.05_{\pm 0.013}$	$0.08_{\pm 0.028}$

M: ground-truth factual mediator; $M_{A'}$: ground-truth counterfactual mediator; $\hat{M}_{A'}$: generated counterfactual mediator

We further give explanations for why our GAN does not copy the factual values but learns the counterfactual values (due to its custom design!). It is true that during training, we cannot directly learn the counterfactual mediators in a supervised way, as they are unobservable. Instead, we can only leverage the reconstruction loss on the factual mediators. The reason why we can still learn the correct counterfactual mediators is due to the adversarial training process of the deterministic generator. By training the discriminator to differentiate between factual and generated counterfactual mediators, the deterministic generator is guided to learn the correct counterfactual distribution.

Importantly, there are three important arguments for why our method does not simply reproduce the factual mediators but actually learns the counterfactual mediators even though they are not observed.

Intuitively: The input of the discriminator D contains the counterfactual and the factual, and the order of them is intentionally randomized. Suppose, hypothetically, that the factual always comes in the first place, then it is easy to distinguish. However, in the design of our framework, this is not the case. In our framework, the data are intentionally shuffled so that factual and counterfactual positions are random.

Technical reason: If the deterministic generator G were to just copy the factual values of the mediator M and output a trivial solution, it would be hard for the discriminator D to distinguish, implying that the loss of D would be large. We would observe mode collapse during training, which we did not observe, and thus provides support for our argument.

Theoretical reason: We provide theoretical proof in our Proposition 1. Therein, we show theoretically that our deterministic generator consistently estimates the counterfactual distribution of the mediator $\mathbb{P}(M_{a'} \mid X = x, A = a, M = m)$. For each specific input data $(X = x, A = a, M = m)$, we then generate the counterfactual mediator from the distribution $\mathbb{P}(M_{a'} \mid X = x, A = a, M = m)$. Hence, we offer theoretical proof that we learn counterfactual fairness correctly.

Appendix K. Computational efficiency

We initially used the approach based on several GANs to ensure that the assumptions of our theoretical foundation were met. However, this was merely done for theoretical reasons, but not for better performance. More specifically, we train several GANs to ensure the worst-case counterfactual fairness is aligned with our theory. This gives the upper bound of the extent to which counterfactual fairness is fulfilled in the prediction model, which is needed for our theoretical guarantees. Below, we demonstrate that a single GAN is sufficient for state-of-the-art performance.

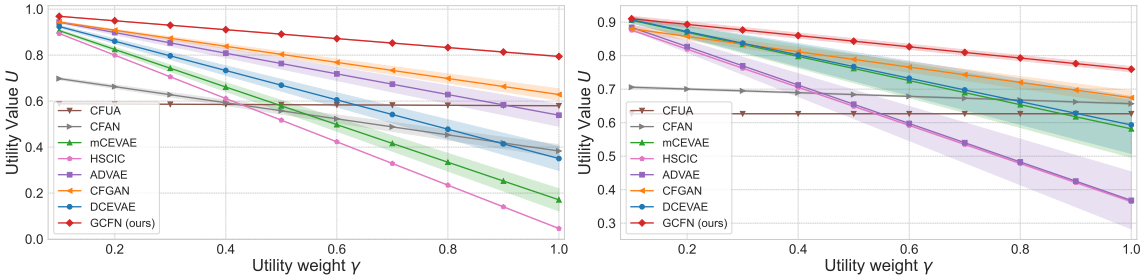

Figure 15: Results for semi-synthetic datasets with a single GAN. A larger utility is better. Shown: mean $\pm$ std over 5 runs.

We conducted the experiments on the LSAC dataset. The results are shown in Figure 15. We can see that our method, based on a single GAN, still gives good results and outperforms all baselines. This demonstrates the scalability of our method. In fact, in practice, the scalability of our proposed method can be simplified using a single GAN. Therefore, we recommend using a single GAN for empirical use and multiple for when theoretical guarantees are additionally needed.

Appendix L. Results for real-world datasets

We now demonstrate the applicability of our method to real-world data. Since ground-truth counterfactuals are unobservable for real-world data, we refrain from benchmarking, but, instead, we now provide additional insights to offer a better understanding of our method.

L.0.1. RESULTS FOR UCI ADULT DATASET

Setting: We use UCI Adult (Asuncion and Newman, 2007) to predict if individuals earn a certain salary but where *gender* is a sensitive attribute. Further details are in Appendix G.

Insights: To better understand the role of our counterfactual mediator regularization, we trained prediction models both with and without applying $\mathcal{R}_{\text{cm}}$. Our primary focus is to show the shifts in the distribution of the predicted target variable (salary) across the sensitive attribute (gender). The corresponding density plots are in Fig. 16. One would expect the distributions for males and females to be more similar if the prediction is fairer. However, we do not see such a tendency for a prediction model without our counterfactual mediator regularization. In contrast, when our counterfactual mediator regularization is used, both distributions are fairly similar as desired. Further visualizations are in Appendix I.

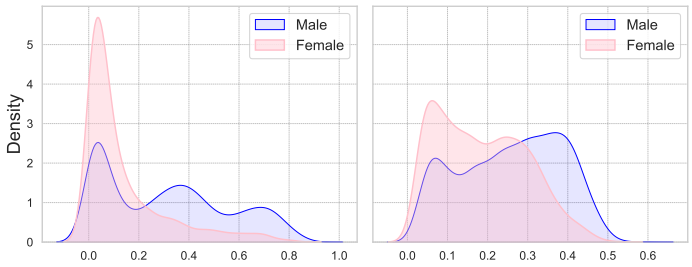

Figure 16: Density of the predicted target variable (salary) across male vs. female. Left: w/o our $\mathcal{R}_{\text{cm}}$. Right: w/ our $\mathcal{R}_{\text{cm}}$.

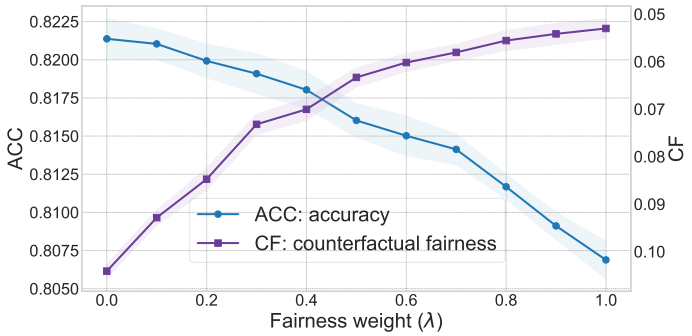

Figure 17: Trade-off between accuracy (ACC) and counterfactual fairness (CF) across different λ for a semi-synthetic data. ACC: the higher (↑) the better. CF: the lower (↓) the better.

Accuracy and fairness trade-off: We vary the fairness weight λ from 0 to 1 to see the trade-off between prediction performance and the level of counterfactual fairness. Since the ground-truth counterfactual is not available for the real-world dataset, we use the generated counterfactual to measure counterfactual fairness on the test dataset. The results are in Fig. 17. In line with our expectations, we see that larger values for λ lead the predictions to be more strict w.r.t counterfactual

fairness, while lower values allow the predictions to have greater accuracy. Hence, the fairness weight λ offers flexibility to decision-makers, so that they can tailor our method to the fairness needs in practice.

L.0.2. RESULTS ON COMPAS DATASET

Setting: We use the COMPAS dataset (Angwin et al., 2016) to predict recidivism risk of criminals and where *race* is a sensitive attribute. The dataset also has a COMPAS score for that purpose, yet it was revealed to have racial biases (Angwin et al., 2016). In particular, black defendants were frequently overestimated in their risk of recidivism. Motivated by this finding, we focus our efforts on reducing such racial biases. Further details about the setting are in Appendix G.

Table 3: Comparison of predictions against actual reoffenses.

Method	ACC	PPV	FPR	FNR
COMPAS	0.6644	0.6874	0.4198	0.2689
GCFN (ours)	0.6753	0.7143	0.3519	0.3032

ACC (accuracy); PPV (positive predictive value);
FPR (false positive rate); FNR (false negative rate).

Insights: We first show how our method adds more fairness to real-world applications. For this, we compare the recidivism predictions from the criminal justice process against the actual reoffenses two years later. Specifically, we compute (i) the accuracy of the official COMPAS score in predicting reoffenses and (ii) the accuracy of our GCFN in predicting the outcomes. The results are in Table 3. We see that our GCFN has a better accuracy. More important is the false positive rate (FPR) for black defendants, which measures how often black defendants are assessed at high risk, even though they do not recidivate. Our GCFN reduces the FPR of black defendants from 41.98% to 35.19%. In sum, our method can effectively decrease the bias towards black defendants.

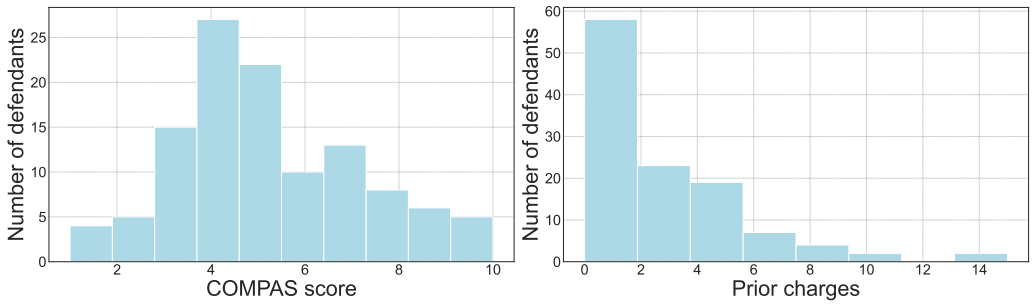

Figure 18: Distribution of black defendants that are treated differently using our GCFN. Left: COMPAS score. Right: Prior charges.

We now provide insights at the defendant level to better understand how black defendants are treated differently by the COMPAS score vs. our GCFN. Fig. 18 shows the number of such different treatments across different characteristics of the defendants. (1) Our GCFN makes oftentimes different predictions for black defendants with a medium COMPAS score of around 4 and 5. However, the predictions for black defendants with a very high or low COMPAS score are similar, potentially

because these are 'clear-cut' cases. (2) Our method arrives at significantly different predictions for patients with low prior charges. This is expected as the COMPAS score overestimates the risk and is known to be biased (Angwin et al., 2016). Further insights are in the Appendix I.

To exemplify the above, Fig. 19 shows two defendants from the data. Both primarily vary in their race (black vs. white) and their number of prior charges (2 vs. 7). Interestingly, the COMPAS score coincides with race, while our method makes predictions that correspond to the prior charges.

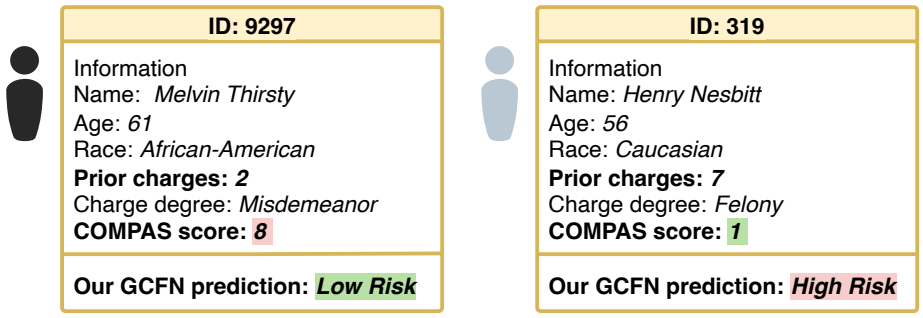

Figure 19: Examples of how defendants are treated differently by the COMPAS score vs. our GCFN.