Counterfactual Fairness with Disentangled Causal Effect Variational Autoencoder

Hyemi Kim[1], **Seungjae Shin**[1], **JoonHo Jang**[1], **Kyungwoo Song**[2], **Weonyoung Joo**[1],
Wanmo Kang[3] and **Il-Chul Moon**[1]

[1]Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST)
[2]Department of AI, University of Seoul, [3]Department of Mathematical Sciences, KAIST
{khm0308, tmdwo0910, adkto8093}@kaist.ac.kr, kyungwoo.song@uos.ac.kr, {es345, wanmo.kang, icmoon}@kaist.ac.kr

Abstract

The problem of fair classification can be mollified if we develop a method to remove the embedded sensitive information from the classification features. This line of separating the sensitive information is developed through the causal inference, and the causal inference enables the counterfactual generations to contrast the what-if case of the opposite sensitive attribute. Along with this separation with the causality, a frequent assumption in the deep latent causal model defines a single latent variable to absorb the entire exogenous uncertainty of the causal graph. However, we claim that such structure cannot distinguish the 1) information caused by the intervention (i.e., sensitive variable) and 2) information correlated with the intervention from the data. Therefore, this paper proposes Disentangled Causal Effect Variational Autoencoder (DCEVAE) to resolve this limitation by disentangling the exogenous uncertainty into two latent variables: either 1) independent to interventions or 2) correlated to interventions without causality. Particularly, our disentangling approach preserves the latent variable correlated to interventions in generating counterfactual examples. We show that our method estimates the total effect and the counterfactual effect without a complete causal graph. By adding a fairness regularization, DCEVAE generates a counterfactual fair dataset while losing less original information. Also, DCEVAE generates natural counterfactual images by only flipping sensitive information. Additionally, we theoretically show the differences in the covariance structures of DCEVAE and prior works from the perspective of the latent disentanglement.

Introduction

Machine learning has penetrated our lives so deep, and its fairness and societal utilization have become a growing concern in our society (Aleo and Svirsky 2008; Kim, Ghorbani, and Zou 2019). The incident of *COMPAS* (Brennan, Dieterich, and Ehret 2009) shows that the learning model can be a source of unfairness in our judicial system by discriminating people by race. Given that the learners become unfair only because of training data (Hardt, Price, and Srebro 2016), we ask the question of whether it is feasible to correct its unfairness from the data or not. Particularly, our concept of unfairness comes from our societal principle on equal treatments across races, gender, religion, etc., a.k.a.
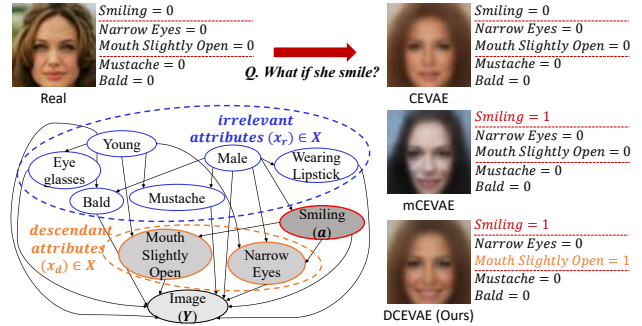
Figure 1: Counterfactual examples have a set of information attributes that are either 1) maintained or 2) altered when the intervention variable, a, is altered. For example, a counterfactually generated image for $a_{Smiling} = 0$ should be labeled as $a_{Smiling} = 1$, and such change may cause the subsequent change on descendant attributes of a, x_d (i.e. *Mouth Slightly Open* and *Narrow Eyes*) and maintaining the other attributes intact, x_r. CEVAE fails to change x_d like *Mouth Slightly Open*, while DCEVAE alters it. mCEVAE fails to maintain attributes which is not caused by a, like hair color for the counterfactual image, but DCEVAE maintains them.

sensitive variables, without prejudice. Then, the key question on the machine learning research becomes whether we can separate such prejudices embedded in the data algorithmically or not.

Considering the prejudice by the sensitive variable, the causal inference is an interesting tool to separate the factors contributing to decision-making. The objective of causal inference is learning the causal effect of an intervention variable, a, on individual features, x, and an outcome, y. Here, if we regard the intervention variable in causality as the sensitive variable in decision-making, the learning fairness can be formulated as the causal inference task (Zhang, Wu, and Wu 2018; Chiappa 2019; Wu, Zhang, and Wu 2019; Kilbertus et al. 2017). For example, a causal model estimates the effect of sensitive variables, such as race and gender, on an admission result (Kusner et al. 2017). Another study shows a causal model enables a fair clinical risk prediction (Pfohl et al. 2019). If we focus on modeling the exogenous uncertainty with Variational Autoencoder (VAE) (Louizos et al.

2017; Pfohl et al. 2019), it has been a common practice to introduce a single latent variable to reflect all exogenous uncertainty.

We separate different causal effects into multiple latent variables, so the diverse aspects of an intervention, features, and an outcome can be related in complex causal graphs. Subsequently, this separation of causal effects by factors enables complex counterfactual example generations because we can only intervene in the sensitive variables by leaving other variables intact. This counterfactual generation becomes our barometer in how fair a learning model is. If a model is fair, the model should result in the same classification for both original and counterfactual instances with an altered sensitive variable.

This paper starts by claiming the limitation of modeling the exogenous uncertainty with a single latent variable, and this paper develops a disentangling structure, or Disentamgled Causal Effect VAE (DCEVAE), for counterfactual generations to relax the limitation. Unlike the previous approaches with a single latent variable to model all features (Shalit, Johansson, and Sontag 2017; Louizos et al. 2017; Pfohl et al. 2019), DCEVAE separates the latent variable to model the exogenous uncertainties either from the intervention or from the feature without the intervention.

As DCEVAE disentangles the uncertainty into two latent variables, DCEVAE has more accurate estimation performances on the total effect and the counterfactual effect compared to Causal VAE models with a single latent variable. DCEVAE added counterfactual fairness regularization to generate counterfactual fair examples with less transformation on the original dataset. Also, DCEVAE generates counterfactual images that do not naturally occur in the dataset, i.e., women with Mustache, through interventions. Finally, we analyze DCEVAE structure from the perspective of linear VAE, and we show DCEVAE is structured to separate the posterior covariance matrix into two parts.

Preliminaries

Counterfactual Fairness Problem Formulation

The final goal of this paper is to provide a counterfactual fair classification method through the latent disentanglement. From this aspect, we start our formulation from the definition of *fairness*. We define A as the sensitive attributes of an individual, which should not be used for discriminative tasks; X as the other observed attributes of individuals; Y as the dependent variable to estimate; and $\hat{Y}$ as the model estimation.

A causal graph specifies $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, \mathbf{P}(u) \rangle$; and $\mathbf{V}$ is the set of endogenous variables, $P(v) := P(V = v) = \sum_{\{u | f_V(V,u)=v\}} P(u)$; and $\mathbf{U}$ is the set of exogenous variables, i.e. the stochastic elements of a variable; and $\mathbf{F}$ is the set of deterministic functions, $V_i = f_{V_i}(PA_{V_i}, U_{V_i})$ with indicating the parents of V_i as PA_{V_i} in a causal graph. With a causal graph, Eq. 1 defines the counterfactual fairness (Kusner et al. 2017).

$$P\big(\hat{Y}_{A \leftarrow a}(U) = y | X = x, A = a\big)$$
$$= P\big(\hat{Y}_{A \leftarrow \neg a}(U) = y | X = x, A = a\big), \quad (1)$$

for all y and any value $\neg a$ attainable by A.

Here, U is the set of exogenous variables. $\hat{Y}_{A \leftarrow a}$ and $\hat{Y}_{A \leftarrow \neg a}$ become two different variations by either a and $\neg a$. The counterfactual fairness asserts that the estimated distributions on $\hat{Y}$s should be identical regardless of the sensitive value, a.

Causality with Variational Autoencoder

Louizos et al. (2017) showed that modeling exogenous variable, U, in a causal graph can be interpreted as an inference task on the latent variables in variational autoencoder (VAE) (Kingma and Welling 2013). The evidence lower bound (ELBO) of VAE, $\mathcal{M}_{ELBO}$, is derived as follows:

$$\log p(x) \geq \mathbb{E}_{q_\phi(u|x)} \left[\log p_\theta(x|u) + \log p(u) - \log q_\phi(u|x) \right]$$
$$=: \mathcal{M}_{ELBO}.$$

Louizos et al. (2017) suggested the modified ELBO in Eq. 2, based on a causal graph. In Causal Effect Variational Autoencoder (CEVAE), a is correlated with all x, and a does not deterministically cause the x. If some of x are caused by a, the causality from a to x will be embedded in u by $q_\phi(u|a, x, y)$ in ELBO. This embedded a in u interrupts the counterfactual generation of $p(y|\neg a, u)$ because the negation only affects a, not the embedded components in u.

$$\mathcal{M}_{CEVAE} = \mathbb{E}_{q(u|a,x,y)}[\log p(u) + \log p(a, x|u)$$
$$+ \log p(y|a, u) - \log q(u|a, x, y)]$$
$$+ \log q(a^*|x^*) + \log q(y^*|x^*, a^*), \quad (2)$$

where x, a^*, y^* being the observed values in the training set.

To compensate this potential problem in CEVAE, modified version of CEVAE (Pfohl et al. 2019), or mCEVAE, assumed that x and y are caused by a and u. $\mathcal{M}_{mCEVAE}$ uses the maximum mean discrepancy (MMD) to regularize the generations to remove the information of a from u, but this MMD regularization removes u components that is simply correlated to a, not caused by a.

$$\mathcal{M}_{mCEVAE} = \mathbb{E}_{q(u|a,x)}\big[\lambda_x \log p(x|a, u) + \lambda_y \log p(y|a, u)\big]$$
$$+ \lambda_1 D_{MMD}\big(q_\phi(u)||p(u)\big)$$
$$+ \lambda_2 \sum_{a_k \in \mathcal{A}} D_{MMD}\big(q_\phi(u|a = a_k)||p(u)\big), \quad (3)$$

where $\mathcal{A}$ is a discrete indicator for a sensitive attribute.

We hypothesize that the accurate counterfactual generation lies in the middle of these two models. We define $x_d \subset x$ is a subset of features caused by a whereas $x_r \subset x$ is the other subset of irrelevant features to the intervention. Similarly, we define the exogenous variables of x_r and x_d to be u_r and u_d, respectively. When the counterfactual generation is required, the intervention on a should be imposed on x_d, and x_r should be maintained. This disentanglement is the fundamental motivation of our model, DCEVAE.

Also, CEVAE separates the decoder network into two alternative functions: $f_{a=0}$ and $f_{a=1}$. Therefore, either decoder tends to be updated according to the observed sensitive variable. For example, i_{th} data $(x_i, a_i = 1, y_i)$ is utilized to learn the parameters of $f_{a=1}$. When CEVAE makes
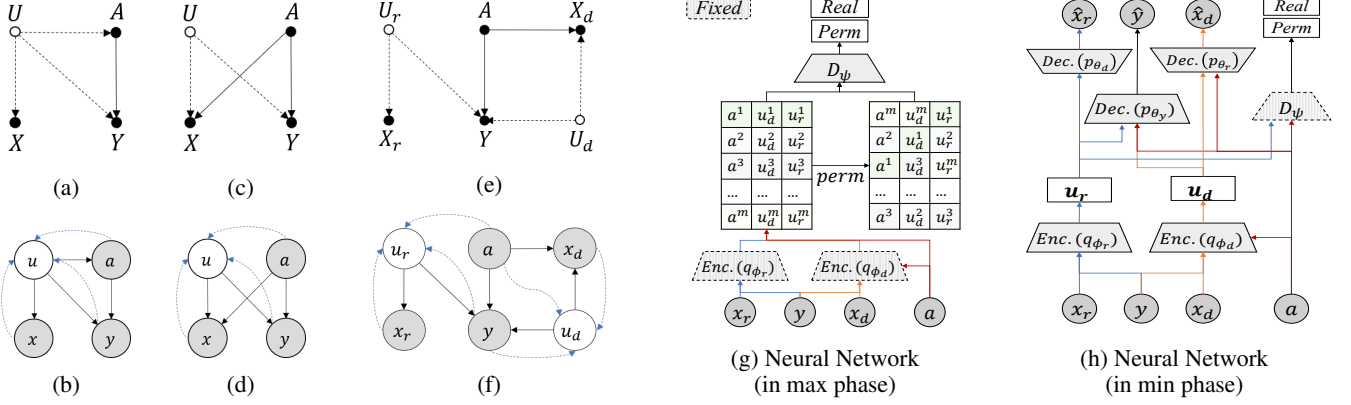
Figure 2: (a,c,e) Causal graph of CEVAE, mCEVAE, and DCEVAE. If a domain and its causal graph is given, the endogenous variables, x, of the given causal graph is mapped to either x_r or x_d; (b,d,f) Bayesian network corresponding to the causal graph of CEVAE, mCEVAE, and DCEVAE. Arrows with solid lines denote generative process, and arrows with dashed lines denote inference process; (g) Neural network structure of DCEVAE in max phase ($\max \mathcal{M}_D$) (h) Neural network structure of DCEVAE in min phase ($\min \mathcal{L}_{DCEVAE}$).

the counterfactual samples, the latent values from a data instance with a are propagated to the decoder $f_{\neg a}$. However, the lack of examples of $\neg a$ in training set can cause inaccuracy of $f_{\neg a}$. To resolve this issue, mCEVAE regularizes latent variables $u_{a=0}$ and $u_{a=1}$ to be similar by using MMD.

Disentanglement with Total Correlation

A latent variable, u, is considered to be disentangled if each dimension of u is independent of each other. When we minimize the total correlation of u (TC), the latent variable is disentangled (Kim and Mnih 2018). Here, $TC = KL\big(q(u)||\Pi_{j=1}^d q(u_j)\big)$, where d is a dimension of u.

Methodology

Causal Structure of Disentangled Causal Effect Variational Autoencoder

We design the structure of DCEVAE from the causal graph in Figure 2e. The causal graph specifies the separated causalities from the sensitive variable, a, to the feature variables, x. Hence, x is divided into the feature variables caused by sensitive information, x_d; and the other feature variables, x_r. This paper assumes that the causal graph gives the attribute association to either x_d or x_r from the domain. For instance, in Figure 1, if we regard $Male$ as a sensitive variable because $Male$ indicates the gender, x_d becomes the set of its descendant variables, $x_d = \{Mustache, Smiling, ...\}$ in the domain causal graph, and x_r is set to be the complementary set of x_d in causal graph variables.

This separation also introduces two corresponding exogenous variables: u_d and u_r. As u_d and u_r are exogenous variables in Figure 2e where we assume that u_r and u_d are disentangled. Also, we assumed that x_d are affected by a, not correlated with a; so u_d needs to be disentangled with a because x_d will be caused by a as an endogenous factor. On the other hand, u_r, which causes x_r, may hold the correlated information of a. Therefore, we did not disentangled u_r with

a since x_r is correlated with a.

The usual set up in potential outcome framework is intervention a precedes outcome y, and all features x are precedes a. However, x being preceded to a is a strong assumption for the real-world case. The more general case is some of x occur before a, the exogenous $u_r \rightarrow x_r$, and the rest of x come after a, the exogenous u_d and a will have a common child x_d. For instance, let us assume that there is a woman (a: *Gender*) who went to a women's only school (x_d: *School*). When we compute counterfactual value by intervening $a = Male$, then the value of school will be changed. However, this person's birth year (x_r), which is not a descendant of gender, will not be changed. This consideration enables a more general causal graph ordering to be incorporated for counterfactual generation based on VAEs.

Bayesian Network of Disentangled Causal Effect Variational Autoencoder

The causal graph in Figure 2e translates to a Bayesian network in Figure 2f. Two exogenous variables, u_r and u_d, in the causal graph are translated into two latent variables in the Bayesian network. This Bayesian network corresponds to the neural network in Figure 2h. The neural network consists of an inference structure on two latent random variables (q_{ϕ_r} and q_{ϕ_d}), and the neural network also includes a structure for disentanglement, as D_ψ. The objective function of DCEVAE, Eq 4, is devised to satisfy the above model structure. The latent variables are inferred by the optimization of $\mathcal{M}_{ELBO}$. The disentanglement of the latent variables is resolved by reducing the total correlation, $\mathcal{L}_{TC}$. In the computation of the total correlation, we use the discriminator D_ψ from $\mathcal{M}_D$, so we add an optimization on $\mathcal{M}_D$ to our objective function, as well. Eventually, our objective becomes the min-max structure to correspond to the counterfactual generation and the latent disentanglement.

$$min_{\theta,\phi} \ \mathcal{L}_{DCEVAE} := -\mathcal{M}_{ELBO} + \beta_{tc}\mathcal{L}_{TC}$$
$$max_\psi \ \mathcal{M}_D \qquad (4)$$

Counterfactual Inference and Counterfactual Example Generation

Besides the model structure, a causal graph requires the inference on the exogenous variable to estimate the causal effect, including the total effect and the counterfactual effect. We match two types of exogenous variables, u_d and u_r, in the causal graph to two corresponding latent variables of DCEVAE. According to the counterfactual inference (Pearl 2009), a counterfactual y can be inferred in following steps:

1. **Abduction** Infer the distribution of u_d and u_r with the encoder network of DCEVAE: $q(u_d|a, x_d, y)$ and $q(u_r|a, x_r, y)$.

2. **Action** Substitue A with $\neg a$.

3. **Prediction** Compute the probability of counterfactual y, with the decoder network of DCEVAE, $p(y|\neg a, u_d, u_r)$.

In the third step, prediction, x_d could be potentially changed for intervention $\neg a$, but x_r maintain its value because we consider causal ordering as $u_r \rightarrow x_r$ and $(a, u_d) \rightarrow x_d$. In Appendix 3.1, we provide identification of our case.

Evidence Lower Bound of DCEVAE

We propose the ELBO to disentangle u_d and u_r by following the Bayesian network structure, Figure 2f. We assume that x_d and u_r is independent given u_d; and x_r and (a, u_d) are independent given u_r, as well. Then, the decoder distribution, $p_\theta(x_d, x_r, y, u_d, u_r|a)$, can be factorized as the below:

$$p_\theta(x_d, x_r, y, u_d, u_r|a) = p(u_d)p(u_r)p_\theta(x_d|a, u_d)$$
$$\times p_\theta(x_r|u_r)p_\theta(y|a, u_d, u_r) \quad (5)$$

Also, we assume that the posterior, $q_\phi(u_d, u_r|a, x_d, x_r, y)$, can be factorized as the Eq. 6.

$$q_\phi(u_d, u_r|a, x_d, x_r, y) = q_\phi(u_d|a, x_d, y)q_\phi(u_r|a, x_r, y) \quad (6)$$

Given an approximate posterior $q_\phi(u_d, u_r|a, x_d, x_r, y)$, we obtain the variational lower bound as Eq. 7.

$$\log p_\theta(x_d, x_r, y|a)$$
$$\geq \mathbb{E}_{q_\phi(u_d|a, x_d, y)q_\phi(u_r|a, x_r, y)}\big[\log p_\theta(y|a, u_d, u_r)\big]$$
$$+ \mathbb{E}_{q_\phi(u_d|a, x_d, y)}\big[\log p_\theta(x_d|a, u_d)\big]$$
$$+ \mathbb{E}_{q_\phi(u_r|a, x_r, y)}\big[\log p_\theta(x_r|u_r)\big]$$
$$+ KL\big(q_\phi(u_d|a, x_d, y)||p(u_d)\big) + KL\big(q_\phi(u_r|a, x_r, y)||p(u_r)\big)$$
$$=: \mathcal{M}_{ELBO} \quad (7)$$

In practice, we use the neural network layers to infer the parameters of a Gaussian distribution over the joint space of u_d and u_r. To obtain the posterior distribution of q_ϕ, $p(u_d)$, and $p(u_r)$ are the prior distributions following the Gaussian distribution, and we utilize the reparametrization, accordingly. The below is the encoder structure for u_d and u_r.

$$p(u_d) = \mathcal{N}(u_d|0, I)f(; \quad q_\phi(u_d|a, x_d, y) = \mathcal{N}(\bar{\mu}_d, \bar{\sigma}_d^2 I);$$
$$p(u_r) = \mathcal{N}(u_r|0, I); \quad q_\phi(u_r|a, x_r, y) = \mathcal{N}(\bar{\mu}_r, \bar{\sigma}_r^2 I);$$
$$\bar{\mu}_d = g_d^\mu(a, x_d, y); \quad \bar{\sigma}_d = g_d^\sigma(a, x_d, y); \quad (8)$$
$$\bar{\mu}_r = g_r^\mu(a, x_r, y); \quad \bar{\sigma}_r = g_r^\sigma(a, x_r, y),$$

With decoder p_θ, DCEVAE provides the generative process for a, x_d, x_r, and y. Since a, x, and y are different in nature, we differentiate their distributions. We assume that if x are continuous variables, they follow the Gaussian distribution. We let y be the binary variable of the Bernoulli trial. x_d and y are influenced by a. On the other hand, x_r is only determined by u_r and not determined by a. The decoder structure for a factual data instance is specified as below:

$$p_{\theta_d}(x_d|a, u_d, y) = \mathcal{N}(\mu = \hat{\mu}_d, \sigma^2 = \hat{\sigma}_d^2 I);$$
$$p_{\theta_r}(x_r|u_r) = \mathcal{N}(\mu = \hat{\mu}_r, \sigma^2 = \hat{\sigma}_r^2 I); \quad (9)$$
$$\hat{\mu}_r, \hat{\sigma}_r^2 = f_r^\mu(u_r), f_r^\sigma(u_r)$$
$$\hat{\mu}_d, \hat{\sigma}_d^2 = f_d^\mu(a, u_d), f_d^\sigma(a, u_d)$$
$$p_{\theta_y}(y|a, u_d, u_r) = Bern(\pi = \hat{\pi}_y); \quad \hat{\pi}_y = f_y(a, u_d, u_r)$$

For counterfactual data generation, we use the same encoder and decoder structure. The only difference is using $\neg a$ for a decoder output. As mentioned before, a counterfactual x_r is same with the factual x_r, while a counterfactual x_d and y are influenced by the counterfactual $\neg a$.

Disentanglement Loss of DCEVAE

Up to this point, we treat each pair (a, u_d) and (u_d, u_r) to be disentangled. This assumption minimize $\mathcal{L}_{TC} = KL\big(q(a, u_r, u_d)||q(a, u_r)q(u_d)\big)$. This KL divergence is intractable because both $q(a, u_d, u_r)$ and $q(a, u_r)q(u_d)$ are conditioned on a, x_d, and x_r. Therefore, we take an alternative approach adopted in FactorVAE (Kim and Mnih 2018). Algorithm 1 in Appendix 3.4 specifies the sampling from $q(a, u_r, u_d)$ and $q(a, u_r)q(u_d)$ under the conditions, and we apply the permutation to minimize $\mathcal{L}_{TC}$. $\mathcal{L}_{TC}$ is related to the discriminator D_ψ by the *density-ratio trick*, which is approximated by a neural network. The output of $D_\psi([a, u_d, u_r])$ estimates the probability when the density function takes a sampled input from $q(a, u_d, u_r)$, rather than from $q(a, u_r)q(u_d)$. $\mathcal{L}_{TC}$ is expressed by D_ψ as in Eq. 10.

$$\mathcal{L}_{TC} = KL\big(q(a, u_d, u_r)||q(a, u_r)q(u_d)\big)$$
$$\approx \mathbb{E}_{q(a, u_d, u_r)}\left[\log \frac{D_\psi(a, u_d, u_r)}{1 - D_\psi(a, u_d, u_r)}\right]. \quad (10)$$

For training the network D_ψ, we should maximize $\mathcal{M}_{D_\psi}$.

$$\mathcal{M}_{D_\psi} = \mathbb{E}_{q(a, u_d, u_r)}\big[\log(D_\psi([a, u_d, u_r]))\big]$$
$$+ \mathbb{E}_{q(a, u_r)q(u_d)}\big[\log 1 - D_\psi([a, u_d, u_r])\big]. \quad (11)$$

Appendix 3.4 provides a whole algorithm of DCEVAE including the minimization of $\mathcal{L}_{DCEVAE}$ and the maximization phase of $\mathcal{M}_D$.

Theoretic Analysis on Covariance Structure

Eq. 12 defines the posterior distribution of the latent variable, and we derive the stationary point of Σ. (Lucas et al. 2019).

$$q(u|a, x, y) = \mathcal{N}\big(V_u([a, x, y] - \mu), \Sigma\big) \quad (12)$$

Here, u is $[u_r, u_d]$, and Σ is a covariance matrix of the joint distribution of latent variables. $\bar{\Sigma}$ is a covariance matrix of the permuted u by discriminator D_ψ.

If $[u_r, u_d]$ is well disentangled, Σ should show two block diagonal matrices corresponding to u_r and u_d. Proposition 1 provides theoretically analysis of the disentanglement.

Proposition 1. *Let Σ^* be the stationary point of the $\mathcal{L}_{DCEVAE}$. For a linear DCEVAE, Σ^* has the form:*

$$\Sigma^* = \left\{ \frac{1}{1+\beta} \left(\frac{1}{\sigma^2} W_r^T M_r^T W_r + \frac{1}{\sigma^2} W_d^T M_d^T W_d \right. \right. \tag{13}$$

$$\left. \left. + diag(\frac{1}{\sigma^2} W_y^T W_y) + I + \beta (\bar{\Sigma}^{-1})^T \right) \right\}^{-1}, \text{ where}$$

$$M_r = \begin{pmatrix} I_{n\times n} & 0_{n\times m} \\ 0_{m\times n} & 0_{m\times m} \end{pmatrix}, M_d = \begin{pmatrix} 0_{n\times n} & 0_{n\times m} \\ 0_{m\times n} & I_{m\times m} \end{pmatrix}$$

with $n = |u_r|$ and $m = |u_d|$.

Proof. *see Appendix 1.1 for the proof.*

We observed that u_r and u_d has their distinct covariance blocks from the real dataset due to the masking effect of M_r and M_d. This theoretically shows the disentangling effects of DCEVAE. Despite the clear disentangling effect from the masks, the off-diagonal covariance can be feasible by $\bar{\Sigma}$, so we provide Corollary 1. Also, the part of covariance, $M_d^T \bar{\Sigma}$, is constructed to be independent to a by the Total Correlation loss, $\mathcal{L}_{TC}$, which enforces u_d to be independent to a.

Corollary 1. *As $\beta \to \infty$, Σ^* becomes a covariance matrix with two blocks on diagonal.*

Proof. $\beta \to \infty \Rightarrow \frac{1}{1+\beta} \to 0, \frac{\beta}{1+\beta} \to 1 \Rightarrow \Sigma^* \to \bar{\Sigma}$. *Also, $\bar{\Sigma}$ is designed to permute u_d and u_r by following Line 8-11, Algorithm 1, Appendix 3.4; $cor(u_d, u_r) \to 0$. Therefore, Σ^* has two blocks of u_d and u_r dimensions.*

Figure 3c and 3d contrast the covariance structure of DCE-VAE to the CEVAE. There is no disentanglement effect on the covariance of CEVAE, so it cannot distinguish the correlated latent variable from the caused ones. Theoretic analyses on the covariance of CEVAE are in Appendix 1.2.

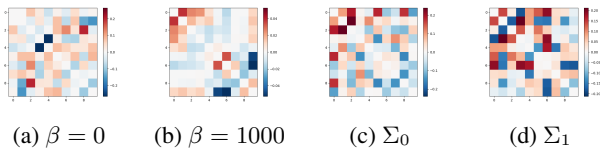(a) $\beta = 0$ (b) $\beta = 1000$ (c) Σ_0 (d) Σ_1

Figure 3: (a,b) The covariance matrix of sampled latent values from DCEVAE with $\beta = 0$ and 1000. The dimension of u_d is five, and so is u_r. (c,d) Each covariance matrix Σ_0 and Σ_1 of sampled latent values from each decoder of CEVAE.

Applications

Causal Fair Classification The task of causal fairness requires estimating y with the minimized influence of a. When we assume $p(\hat{y}|a, u_{d,i}, u_{r,i}) - p(\hat{y}|\neg a, u_{d,i}, u_{r,i}) = 0$ for $u_{d,i}$ and $u_{r,i}$ from i-th data instance; we say that the counterfactual fairness is satisfied for the data instance. Therefore, we alter the objective function of DCEVAE by adding a regularization $\mathcal{L}_f$ as the below:

$$\min \mathcal{L}_{fair} = \mathcal{L}_{DCEVAE} + \beta_f \mathcal{L}_f, \ \max \mathcal{M}_D \text{ where} \tag{14}$$

$$\mathcal{L}_f = \mathbb{E}_{q(u_d, u_r|a, x_d, x_r)} \left[||p_\theta(y|a, u_d, u_r) - p_\theta(y|\neg a, u_d, u_r)||_2 \right]$$

After optimizing Eq. 14, we train a classifier, such as a logistic regression (LR), with the pairs of $\hat{y}$ with $\hat{x}$ through the decoder, $p_\theta(x_d, x_r, y|a, u_d, u_r)$ and $p_\theta(x_d, x_r, y|\neg a, u_d, u_r)$. The test procedure utilizes the raw input of the testing feature a and x without any modifications.

Counterfactual Image Generation The counterfactual image generation task allocates y to be the image and x to be the labels which describe the image. a is the label which we want to intervene on. Unlike the fairness dataset, we modified the encoder, $q_\phi(u_d, u_r|a, y)$, because the information of x is already embedded on an image, y. The counterfactual images are sampled from the decoder, $p_\theta(y|\neg a, u_d, u_r)$, while u_d and u_r are obtained from the encoder.

Experiments

Datasets and Baselines

Appendix 4.1 provides the dataset details; and Appendix 4.2 enumerates the causal graphs and information of (x_d, x_r).

Causal Estimation and Fair Classification We use the UCI Adult income dataset (Dua and Graff 2017) for causal estimation and fair classification tasks. We treat *gender* as a sensitive varible (or intervention) a; *income* as the outcome y; *race*, *age*, and *native country* as x_r; and other variables as x_d. We benchmark the causal effect estimation on five baselines: CausalGAN (Kocaoglu et al. 2017), CausalGAN-Incomplete (CausalGAN-IC), conditional VAE (CVAE) (Sohn, Lee, and Yan 2015), CEVAE, and mCE-VAE. CausalGAN-IC has the same generator and discriminator structures as CausalGAN, but CausalGAN-IC used the same causal graph DCEVAE used. In fairness task, we additionally use Unawareness (CF-UA), Additive Noise (CF-AN) (Kusner et al. 2017), and CFGAN (Xu et al. 2019) as baselines.

Counterfactual Image Generation We use the CelebA dataset (Liu et al. 2018) for the counterfactual image generation. We treat *Mustache* as an intervention attribute a; an image as y; and the other attributes as x. Appendix 5.1 provide a similar experiment with *Smiling* as an intervention attribute. We seperate x_d and x_r as the assumed causal graph in Figure 1. We consider the following baselines for counterfactual image generations: CVAE , CEVAE, mCEVAE, Conditional GAN (CGAN) (Mirza and Osindero 2014) with Wasserstein distance (Arjovsky, Chintala, and Bottou 2017) (denoted as CWGAN), CausalGAN, and CausalGAN-IC.

Evaluation Metrics

We evaluate the performance on the causal estimation and fairness task with following metrics used in CFGAN: **Total effect**, $TE(a_1, a_0) = P(y_{a_1}) - P(y_{a_0})$, measures the change likelihood of A from a_0 to a_1 on Y. **Counterfactual effect**, $CE(a_1, a_0|o) = P(y_{a_1}|o) - P(y_{a_0}|o)$, is the total effect conditioned on the observation, o. A **Logistic Regression (LR)** and a **Support Vector Machine (SVM)** are trained with generated datasets from the model, and their test accuracy is calculated on the original dataset. **Chi square distance** (χ^2) indicates the similarity between the generated and the real datasets (Daliri 2013).

	Total Effect	Counterfactual Effect (CE)				CE error	χ^2 $(\times 10^{-2})$	Accuracy	
		o_{00}	o_{01}	o_{10}	o_{11}			LR	SVM
Real Data	.1936	.1785	.1266	.1293	.2023	-	0	.8158	.8109
CausalGAN	$.1959_{\pm.0140}$	$.1123_{\pm.0231}$	$.1437_{\pm.0128}$	$.1690_{\pm.0144}$	$.2080_{\pm.0163}$	$.1374_{\pm.0267}$	$4.20_{\pm.08}$	$.7805_{\pm.0220}$	$.8126_{\pm.0026}$
CausalGAN-IC	$.2432_{\pm.0163}$	$\underline{.1585}_{\pm.0622}$	$.1912_{\pm.0347}$	$.2267_{\pm.0373}$	$.2556_{\pm.0162}$	$.2689_{\pm.0663}$	$4.65_{\pm.21}$	$.7994_{\pm.0141}$	$\mathbf{.8049}_{\pm.0129}$
CVAE	$.1510_{\pm.0279}$	$.1228_{\pm.0177}$	$\mathbf{.1194}_{\pm.0241}$	$.1347_{\pm.0264}$	$.1452_{\pm.0295}$	$.1524_{\pm.0618}$	$0.25_{\pm.01}$	$.7945_{\pm.0040}$	$.7886_{\pm.0064}$
CEVAE	$.0954_{\pm.0036}$	$.0951_{\pm.0066}$	$.0843_{\pm.0017}$	$.0921_{\pm.0044}$	$.0972_{\pm.0041}$	$.3399_{\pm.1664}$	$\mathbf{0.18}_{\pm.01}$	$.7909_{\pm.0014}$	$.7615_{\pm.0000}$
mCEVAE	$.1480_{\pm.0160}$	$.1296_{\pm.0858}$	$.1166_{\pm.0422}$	$\mathbf{.1340}_{\pm.0431}$	$.1543_{\pm.0164}$	$.1790_{\pm.1190}$	$0.26_{\pm.03}$	$.7877_{\pm.0199}$	$.7884_{\pm.0246}$
DCEVAE (ours)	$\mathbf{.1831}_{\pm.0011}$	$\mathbf{.1871}_{\pm.0086}$	$\underline{.1574}_{\pm.0054}$	$.1673_{\pm.0068}$	$\mathbf{.1878}_{\pm.0010}$	$\mathbf{.0923}_{\pm.0071}$	$0.27_{\pm.01}$	$\mathbf{.8095}_{\pm.0050}$	$\underline{.7995}_{\pm.0003}$

Table 1: The total effect and counterfactual effect of real and generated datasets ($O = \{$race, native country$\}$). CE error is $\sum_{i,j\in\{0,1\}} \left| \frac{o_{ij} - o_{ij}^*}{4} \right|$ with true CE, o^*. The numbers in bold indicates the best performance, and the underlined numbers indicate the second best performance.

We follows MaskGAN (Lee et al. 2019) to evaluate quality of counterfactually generated images: **Semantic-level Evaluation** evaluates generated images by measuring the preservation of the original values of x_r. We trained a classifier with ResNet-18 (He et al. 2016) to examine the accuracy of x_r in generated images. **Distribution-level Evaluation** measures the quality and the diversity of generated images, and we used the Frechet Inception Distance (FID) (Cao et al. 2013). **Identity Preserving Evaluation** evaluates the identity preservation ability, so we conducted a face matching experiment for whole pairs of original and counterfactual images with ArcFace (Deng et al. 2019). We use $1,544$ pairs for *Mustache* and $10,000$ pairs for *Smiling*.

Results

Causal Estimation and Fairness Task

Without the regularization of the fairness $\mathcal{L}_{fair}$, we calculate the total effect (TE) and the counterfactual effect (CE) of our model and baselines in Table 1 for UCI Adult. DCEVAE estimates the total effect and the counterfactual effect close to the original dataset through DCEVAE does not know the exact causal graph structure, unlike CausalGAN. CausalGAN estimates the true TE and CE well, but CausalGAN-IC has lower CE and TE estimation accuracies when the incomplete causal graph is given. CEVAE has lower performance on TE and CE estimations compared to DCEVAE, which is caused by the latent variable in CEVAE with the correlated information from a to x_d. Therefore, CEVAE maintains correlated information even in the counterfactual prediction.

With the regularization of the fairness, $\mathcal{L}_{fair}$, Figure 4 shows the results of fairness tasks for UCI Adult. The training datasets from CFGAN have high accuracy when LR and SVM are selected as classifiers, and CFGAN has a low value of TE and the average of CE. However, when an incomplete graph is given, CFGAN-IC has the low accuracy of LR and SVM, and the high values of TE and CE. In contrast to the low reliability of CFGAN, DCEVAE has comparable accuracies, TE and CE to CFGAN without causal graph structure. Also, the variance of TE and CE from DCEVAE shows the reliability compared to CFGAN.
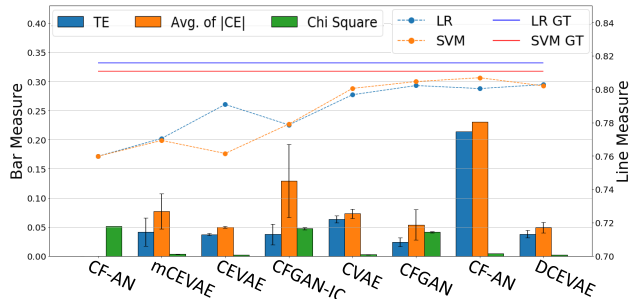

Figure 4: (Left axis) TE, Average of the absolute value of CE, and χ^2 (Chi-Square), (Right axis) Accuracy of LR and SVM For fairly generated dataset. GT is ground truth.

Image Generation Task

We choose *Mustache* and *Smiling* as intervention variables in the CelebA dataset. Our model and baselines generate the counterfactual image by negating the value of the intervention variable. For example, if the real image has *Mustache* = 0, the counterfactual image should have *Mustache* = 1.

This section describes (1) the visualization of counterfactual images from our model and baselines; (2) the analysis of the latent variables from each model; and (3) the quantitative analysis for the image generation task.

Generated Counterfactual Images Figure 5 visualizes the counterfactual images from the VAE based models for the real image. VAE based models can infer the exogenous variables from a specific real image. Counterfactual images from DCEVAE preserve the identity of real images except for the intervention variables. For example, female images on the first row in Figure 5 are kept to be female while the intervened *Mustache* (M) is added. On the contrary, CVAE makes images with *Mustache*, but its gender is altered. CE-VAE creates blurry images because CEVAE separates the decoders by the case of interventions, so a female image is rarely given to training the decoder of $M = 1$.

mCEVAE also fails in generating images because the latent variable loses information as MMD enforces the latent variables of $M = 0$ and $M = 1$ to be overlapped. Also, the

	Model	Target	Attribute clssification accuracy (%)								IP	FID
			WL	MSO	S	ML	E	B	NE	Y		
Real	CVAE	$46.1_{\pm3.9}$	$85.4_{\pm0.4}$	$73.7_{\pm0.7}$	$80.5_{\pm1.2}$	$76.5_{\pm2.6}$	$65.5_{\pm2.9}$	$78.6_{\pm3.3}$	$62.3_{\pm2.9}$	$65.2_{\pm2.8}$	$.26_{\pm.04}$	$187.88_{\pm6.09}$
	CEVAE	$\mathbf{50.4_{\pm0.5}}$	$85.3_{\pm0.2}$	$74.6_{\pm0.2}$	$79.9_{\pm0.5}$	$77.0_{\pm2.1}$	$69.0_{\pm4.3}$	$\mathbf{82.1_{\pm1.1}}$	$65.1_{\pm0.7}$	$64.5_{\pm2.0}$	$.29_{\pm.04}$	$181.29_{\pm8.57}$
	mCEVAE	$44.3_{\pm4.5}$	$85.4_{\pm0.3}$	$73.4_{\pm1.3}$	$79.8_{\pm1.1}$	$76.6_{\pm1.5}$	$59.9_{\pm1.8}$	$79.2_{\pm0.8}$	$63.3_{\pm4.2}$	$65.1_{\pm1.9}$	$.23_{\pm.02}$	$\mathbf{175.30_{\pm1.26}}$
	DCEVAE	$49.7_{\pm0.2}$	$\mathbf{85.6_{\pm0.6}}$	$\mathbf{74.8_{\pm0.4}}$	$\mathbf{81.5_{\pm0.2}}$	$\mathbf{78.9_{\pm0.5}}$	$\mathbf{71.3_{\pm2.3}}$	$\mathbf{82.1_{\pm0.2}}$	$\mathbf{66.3_{\pm0.5}}$	$\mathbf{67.7_{\pm0.9}}$	$\mathbf{.33_{\pm.02}}$	$176.55_{\pm0.78}$
Pair	CVAE	$48.5_{\pm2.0}$	$93.5_{\pm2.2}$	$95.1_{\pm2.7}$	$92.3_{\pm1.9}$	$85.8_{\pm2.4}$	$75.2_{\pm7.0}$	$86.0_{\pm6.1}$	$75.3_{\pm4.2}$	$77.3_{\pm9.2}$	$.42_{\pm0.17}$	$187.88_{\pm6.1}$
	CEVAE	$50.4_{\pm0.3}$	$\mathbf{96.7_{\pm1.4}}$	$\mathbf{95.4_{\pm2.0}}$	$92.8_{\pm3.4}$	$89.0_{\pm4.9}$	$88.3_{\pm1.4}$	$93.8_{\pm1.1}$	$86.4_{\pm3.2}$	$85.6_{\pm3.6}$	$.75_{\pm.07}$	$181.29_{\pm8.57}$
	mCEVAE	$47.5_{\pm2.3}$	$93.7_{\pm1.4}$	$94.7_{\pm3.1}$	$92.1_{\pm1.2}$	$84.7_{\pm2.2}$	$65.8_{\pm2.0}$	$87.1_{\pm3.5}$	$77.9_{\pm8.5}$	$78.7_{\pm4.9}$	$.43_{\pm.21}$	$175.30_{\pm1.26}$
	DCEVAE	$50.1_{\pm0.2}$	$93.8_{\pm1.3}$	$95.1_{\pm1.1}$	$\mathbf{94.7_{\pm0.9}}$	$\mathbf{93.9_{\pm0.7}}$	$\mathbf{92.0_{\pm1.8}}$	$\mathbf{95.1_{\pm0.7}}$	$\mathbf{91.2_{\pm0.8}}$	$\mathbf{88.8_{\pm1.2}}$	$\mathbf{.98_{\pm.01}}$	$176.55_{\pm0.78}$
	CWGAN	$\mathbf{55.9_{\pm1.6}}$	$93.9_{\pm1.0}$	$75.1_{\pm2.2}$	$85.5_{\pm1.0}$	$62.5_{\pm1.2}$	$85.6_{\pm0.9}$	$81.3_{\pm2.3}$	$69.3_{\pm0.9}$	$65.9_{\pm1.2}$	$.11_{\pm.01}$	$\mathbf{106.14_{\pm1.16}}$
	DCGAN	$51.2_{\pm0.4}$	$81.5_{\pm5.7}$	$66.6_{\pm3.5}$	$75.6_{\pm6.5}$	$63.5_{\pm3.6}$	$78.1_{\pm3.8}$	$88.6_{\pm3.3}$	$72.2_{\pm3.6}$	$66.9_{\pm2.8}$	$.13_{\pm.02}$	$141.94_{\pm3.60}$
	BEGAN	$50.9_{\pm1.8}$	$94.1_{\pm3.1}$	$73.8_{\pm6.7}$	$89.0_{\pm5.7}$	$69.2_{\pm8.1}$	$82.4_{\pm2.7}$	$83.6_{\pm9.1}$	$76.4_{\pm3.3}$	$72.5_{\pm6.9}$	$.30_{\pm.35}$	$192.11_{\pm93.46}$
	DCGAN-IC	$50.9_{\pm1.0}$	$80.2_{\pm3.3}$	$62.7_{\pm2.9}$	$71.0_{\pm1.9}$	$63.9_{\pm3.0}$	$75.1_{\pm5.9}$	$89.6_{\pm2.7}$	$72.8_{\pm4.4}$	$68.2_{\pm2.3}$	$.09_{\pm.01}$	$139.80_{\pm2.72}$
	BEGAN-IC	$\underline{51.6_{\pm1.1}}$	$89.9_{\pm1.7}$	$60.6_{\pm1.4}$	$82.1_{\pm3.2}$	$62.2_{\pm1.6}$	$75.9_{\pm3.2}$	$86.2_{\pm3.7}$	$71.8_{\pm2.4}$	$66.4_{\pm2.1}$	$.08_{\pm.02}$	$145.9_{\pm2.89}$

Table 2: (1) Real: comparison between real images and generated counterfactual examples; Pair: comparison between reconstructed images and counterfactual images with the same latent values, (2) Target (*Mustache*) Accuracy, (3) Identity Preserving (IP) scores (threshold= 0.2 for Real, and 0.6 for Pair), (4) The attribute classification accuracy for unaltered labels, (5) FID Score (FID). The numbers in bold indicate the best performance, and the underlined numbers indicate that the second best performance. (Here, M: *Mustache*; WL: *Wearing Lipstick*, MSO: *Mouth Slightly Open*, S: *Smiling*, ML: *Male*, E: *Eyeglasses*, B: *Bald*, NE: *Narrow Eyes*, and Y: *Young*).

joint distribution of $M = 1, Male = 1$ and the joint distribution of $M = 0, Male = 1$ have a large overlapping area, so mCEVAE can make counterfactual images with respect to $M = 1$ for images with *Male* = 1, but not for images with $M = 0$ and *Male* = 0, see Figure 5. It should be noted that we did not include the GAN-based counterfactual generation because the GAN variations cannot produce the counterfactual image matched to a given real-world image.
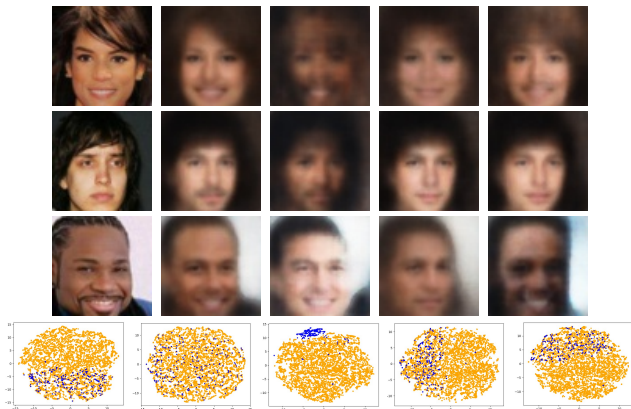

Figure 5: (rows 1-3) Real images and its counterfactual images for *Mustache* from DCEVAE, CVAE, CEVAE, and mCEVAE. (row 4) tSNE (Maaten and Hinton 2008) of latent variables u_r and u_d from DCEVAE; u from CVAE, CEVAE, and mCEVAE. Here, $M = 0$ (yellow) and $M = 1$ (blue).

Quantitative Analysis on Image Generation Task This section shows the quantitative evaluations on generated counterfactual images with label classifier accuracies, Frechet Inception Distance (FID) score, and identity preserving (IP) metrics. Table 2 shows the result of counterfactual generation on *Mustache*. We compare the real and the counterfactual generated images of VAE-based approaches. DCEVAE has the highest attribute classification accuracies for most of x_r, so DCEVAE maintains the attribute of x_r intact. Also, DCEVAE has the highest IP and low FID scores, so the generated images are evaluated to be more natural than the other models. The generation of $Mustache$ is measured by a classifier accuracy, and CEVAE, mCEVAE, and DCEVAE have similar accuracies. We compare reconstructed images and counterfactual images from VAEs and GANs. Among all models, DCEVAE preserves attributes causing *Mustache*, i.e. preserved *Male* attribute.

Conclusion

This paper disentangles the exogenous uncertainty into two latent variables of 1) independent to interventions (u_d), and 2) correlated to interventions without causality (u_r). The disentanglement of latent variables resolves the limitation in previous works, including maintaining causality from the intervention (a) and altering all correlated information for counterfactual instances. We theoretically show that the covariance structure of DCEVAE is disentangled unlike prior works. Our model, DCEVAE, estimates the total effect and the counterfactual effect without a complete causal graph. In experiments, we show that the performance of DCEVAE is comparable with that of other models with and without the complete causal graph for both applications including fair classification and the counterfactual image generation.

Acknowledgements

Ethical Impact

Besides the COMPAS incident (Brennan, Dieterich, and Ehret 2009), governments and corporates utilize the AI-based screening and recommendation systems on a massive scale, and these applications are prone to the fairness question, particularly when the subject individual has minority backgrounds. This paper discusses the triad of 1) the fair classification, 2) the causality-based counterfactual generation, and 3) the latent disentanglement. If we were to maximize the classification accuracy, the proposed method would be irrelevant from such efficiency-oriented perspectives. However, our society always asks what-if questions, i.e., the veil of ignorance by Rawls (Rawls 2009). The limitation of the accuracy can be acceptable in two conditions: 1) the damage to the accuracy performance should be controlled and minimal, and 2) the limitation satisfies the argument that "I would accept the classification result under my altered background." This "altered background" is, in fact, an identical argument to the justice concept suggested by Rawls, which argues designing a taxation concept before determine whether you will be born in either high-income or low-income families. An emerging question is whether or not an AI can come up with a justifiable altered concept on what-if scenarios, so we work on the counterfactual generation to satisfy the fairness concept defined in the above. This counterfactual generation is further elaborated if an AI carefully dissects the context of an individual subject, and this is the disentanglement process when causality and a correlation should be distinguished. Given this series of arguments and necessity, this work plays a vital role in promoting the fairness of deployed AI systems, which are already running without a user's perception of its background operation.

References

Aleo, M.; and Svirsky, P. 2008. Foreclosure fallout: The banking industry's attack on disparate impact race discrimination claims under the fair housing act and the equal credit opportunity act. *BU Pub. Int. LJ* 18: 1.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* .

Brennan, T.; Dieterich, W.; and Ehret, B. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36(1): 21–40.

Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20(3): 413–425.

Chiappa, S. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7801–7808.

Daliri, M. R. 2013. Chi-square distance kernel of the gaits for the diagnosis of Parkinson's disease. *Biomedical Signal Processing and Control* 8(1): 66–70.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. URL http://archive.ics.uci.edu/ml.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kilbertus, N.; Carulla, M. R.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, 656–666.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. *arXiv preprint arXiv:1802.05983* .

Kim, M. P.; Ghorbani, A.; and Zou, J. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 247–254.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2017. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023* .

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4066–4076.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2019. MaskGAN: towards diverse and interactive facial image manipulation. *arXiv preprint arXiv:1907.11922* .

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August* 15: 2018.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, 6446–6456.

Lucas, J.; Tucker, G.; Grosse, R. B.; and Norouzi, M. 2019. Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse. In *Advances in Neural Information Processing Systems*, 9408–9418.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Pearl, J. 2009. *Causality*. Cambridge university press.

Pfohl, S.; Duan, T.; Ding, D. Y.; and Shah, N. H. 2019. Counterfactual Reasoning for Fair Clinical Risk Prediction. *arXiv preprint arXiv:1907.06260* .

Rawls, J. 2009. *A theory of justice*. Harvard university press.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3076–3085. JMLR. org.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, 3483–3491.

Wu, Y.; Zhang, L.; and Wu, X. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, 10–16.

Xu, D.; Wu, Y.; Yuan, S.; Zhang, L.; and Wu, X. 2019. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Zhang, L.; Wu, Y.; and Wu, X. 2018. Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering* 31(11): 2035–2050.