An Information-Theoretic Quantification of Discrimination with Exempt Features

Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, Pulkit Grover
Carnegie Mellon University
sanghamd@andrew.cmu.edu, {vpraveen, piotrm, danupam, pulkit}@cmu.edu

Abstract

The needs of a business (e.g., hiring) may require the use of certain features that are critical in a way that any discrimination arising due to them should be exempted. In this work, we propose a novel information-theoretic decomposition of the total discrimination (in a counterfactual sense) into a non-exempt component, which quantifies the part of the discrimination that cannot be accounted for by the critical features, and an exempt component, which quantifies the remaining discrimination. Our decomposition enables selective removal of the non-exempt component if desired. We arrive at this decomposition through examples and counterexamples that enable us to first obtain a set of desirable properties that any measure of non-exempt discrimination should satisfy. We then demonstrate that our proposed quantification of non-exempt discrimination satisfies all of them. This decomposition leverages a body of work from information theory called Partial Information Decomposition (PID). We also obtain an impossibility result showing that no observational measure of non-exempt discrimination can satisfy all of the desired properties, which leads us to relax our goals and examine alternative observational measures that satisfy only some of these properties. We then perform a case study using one observational measure to show how one might train a model allowing for exemption of discrimination due to critical features.

1 Introduction

As artificial intelligence becomes ubiquitous, it is important to understand whether a machine-learnt model is perpetuating existing biases, and if so, how we can engineer fairness into such a model. The field of fair machine learning (Dwork et al. 2012; Agarwal et al. 2018; Hardt et al. 2016; Calmon et al. 2017; Kamishima et al. 2012; Wang, Ustun, and Calmon 2019; Menon and Williamson 2018; Komiyama and Shimao 2017; Donini et al. 2018; Ghassami, Khodadadian, and Kiyavash 2018; Heidari et al. 2018; Liao et al. 2019; Varshney 2019) provides several measures of fairness, and uses them to reduce discrimination based on *protected attributes*, e.g., as a regularizer during training (Agarwal et al. 2018; Kamishima et al. 2012).

In particular applications, there are some features that are *critical* in a way that they are required to be weighed strongly in the decision *even if* they perpetuate bias, e.g., educational qualification for a job, merit and seniority in deciding salary etc. (Kamiran, Žliobaitė, and Calders 2013). Hence, the discrimination arising due to these features can be exempted. In this work, our goal is to formalize and quantify the *non-exempt discrimination*, i.e., the part of the discrimination that cannot be accounted for by critical features, and selectively remove it if desired.

While such categorization of features is application-dependent and might require domain knowledge and ethical evaluation, such exemptions do exist. E.g., the US Equal Pay Act (US Equal Employment Opportunity Commission) exempts for any difference in salary based on gender that can be explained by merit and seniority. Similarly, the US employment discrimination law (Manley 2009) contains a Bona Fide Occupational Qualification (BFOQ) defense where discrimination based on protected attributes may be exempted if the discrimination is due to a BFOQ reasonably necessary to the normal operation of that particular business, or other reasonable differentials. E.g., fire departments may require firemen to be able to lift a given weight to demonstrate that they will be able to carry fire victims out of a burning building. This feature is therefore allowed to be weighed strongly in hiring even if it is correlated with protected attributes. Similarly, UK employment discrimination law also allows exemptions based on the privacy and decency of the people the employer would be dealing with, e.g., staff in a care home (Dept. of Trade and Industry 2008).

In this work, we assume that the critical features are known (similar to Kamiran, Žliobaitė, and Calders (2013), Kilbertus et al. (2017), Salimi et al. (2019)). Let X_c and X_g denote the critical and the non-critical or general features, respectively. We denote the protected attribute(s) as Z and the model output as $\hat{Y}$. Note that $\hat{Y}$ is a function of the entire feature vector $X = (X_c, X_g)$.

Why should a model use the "general" features at all for prediction if they are not critical? The general features can improve accuracy, or reduce the candidate pool, e.g., if 60% of applicants clear a test but resources are available to interview only 10%. Not using the general features at all may

reduce accuracy or produce a very large candidate pool. Our goal is to use both critical and general features in a way that maximizes accuracy (to the extent possible) while preventing only the non-exempt discrimination.[1]

In this work, our contributions are as follows:

1. Quantification of Non-Exempt Discrimination: As a first step towards this quantification, we propose an information-theoretic quantification of the total discrimination (exempt and non-exempt) that is 0 if and only if the "counterfactual causal influence" (Kusner et al. 2017) is 0, i.e., the model is *counterfactually fair*. Intuitively speaking, we extend the idea of "proxy-use"(Datta et al. 2017) from white-box models to black-box models, where we regard a model as being discriminatory if a virtual component (P) is formed inside the model that has high mutual information about Z (i.e., P is a virtual proxy of Z) and that also causally influences the final output $\hat{Y}$. Interestingly, note that this discrimination may not exhibit itself entirely in $\mathrm{I}(Z;\hat{Y})$, which is the "statistically visible" information about Z in $\hat{Y}$ because of "statistical masking effects," e.g., $\hat{Y} = P + G$ where $G \perp\!\!\!\perp Z$.

Next, we quantify the *non-exempt* part of this discrimination. Our quantification leverages a body of work in information theory called Partial Information Decomposition (PID). We consider examples and thought experiments to arrive at a set of desirable properties that any measure of non-exempt discrimination should satisfy, and then provide a measure that satisfies them (see Theorem 1). First, we require the measure to be 0 if all the features are in the exempt set X_c. Next, it is desirable that the measure be non-zero if $\hat{Y}$ has any "unique" information about Z that is not present in X_c because then that information content is also attributed to X_g. However, because of statistical masking effects, even if this unique information is 0, there may still be non-exempt masked discrimination. Lastly, the measure should not capture false positives, e.g., it should be 0 if such virtual proxies cancel each other such that the final model output has no counterfactual causal influence of Z.

2. Decomposition of Total Discrimination: Next, we propose the decomposition of total discrimination into four non-negative components, namely, exempt and non-exempt visible discrimination and exempt and non-exempt masked discrimination (see Theorem 2).

3. An Impossibility Result: We show that no purely observational measure of non-exempt discrimination can satisfy all our desirable properties (see Theorem 3).

4. Observational Relaxations: Relaxing our requirements, we obtain purely observational measures that satisfy some of the desirable properties (summarized in Table 1) and then use one of them, namely, conditional mutual information, to demonstrate how to selectively reduce non-exempt discrimination in practice through a case study.

Related Work: We are aware that the idea of using conditional mutual information as a metric for non-exempt discrimination has surfaced in another work (Anonymous 2019), where the focus is on conditional debiasing of neural networks using novel estimators. Other observational measures of non-exempt discrimination have also been discussed in Kamiran, Žliobaitė, and Calders (2013), Zafar et al. (2015), Salimi et al. (2019), Corbett-Davies et al. (2017). In this work, our focus is on an axiomatic examination of such measures and their relationship with the concept of *counterfactual fairness*[2] which has not received detailed attention. We also examine and acknowledge the utility and limitations of our observational measures (e.g., see an impossibility result in Theorem 3).

Causal approaches for fairness have been explored in Kusner et al. (2017), Kilbertus et al. (2017), Russell et al. (2017), Chiappa and Gillam (2018), Datta et al. (2017), including impossibility results on purely observational measures (Kilbertus et al. 2017; Datta et al. 2017). The main novelty arises from our adoption of a proxy-use viewpoint for black-box models *that allows for feature exemptions*. The decomposition of total discrimination into exempt and non-exempt components is tricky: one might be tempted to examine specific causal paths from Z to $\hat{Y}$ that pass (or do not pass) through X_c, and deem those influences as the two measures. However, as the PID literature notes, discrimination can also arise from synergistic information (Venkatesh, Dutta, and Grover 2019; Bertschinger et al. 2014; Williams and Beer 2010) about Z in both X_c and X_g, that cannot be attributed to any one of them alone, *i.e.*, $\mathrm{I}(Z;X_c)$ and $\mathrm{I}(Z;X_g)$ may both be 0 but $\mathrm{I}(Z;X_c,X_g)$ may not (see Counterexample 3). Purely causal measures (that do not rely on the PID framework) can attribute such discrimination entirely to X_c. We contend that such synergistic information, if influencing the decision, must be included in the *non-exempt* component of discrimination because, operationally, both X_c and X_g are contributors. We also note that identifying synergy is important: synergy arises frequently in machine-learning (Tax, Mediano, and Shanahan 2017).

In a sense, this work treads a middle ground between two schools of thought, namely, *demographic parity* (Agarwal et al. 2018; Ghassami, Khodadadian, and Kiyavash 2018), which enforces the criterion $Z \perp\!\!\!\perp \hat{Y}$, and *equalized odds* (Hardt et al. 2016; Ghassami, Khodadadian, and Kiyavash 2018), which enforces $Z \perp\!\!\!\perp \hat{Y}|Y$ (directly or through practical relaxations) where Y denotes the true labels of the historic dataset. Our selective quantification of non-exempt discrimination helps address one of the major criticisms against demographic parity, namely, that it can deliberately choose unqualified members from the protected group (Zemel et al. 2013), e.g., by disregarding the critical features if they are correlated with Z. Another strength of our approach is that it does not use the true labels for

[1] Example (inspired by Barocas and Selbst (2016)): To choose a "good" employee, an employer could evaluate standardized test scores and reference letters (human-graded performance reviews). Both features are "job-related" in that they have statistical correlation with the prediction goal and can help improve accuracy. However, test-scores, a critical feature, should be weighed strongly in the decision *even if* biased whereas reference letters may be used only to the extent that they do not discriminate.

[2] Our measure of total (exempt and non-exempt) discrimination is zero if and only if the "counterfactual causal influence" of Z on $\hat{Y}$ is zero (see Lemma 1).

Table 1: Observational Measures (M_{NE}) of Non-Exempt Discrimination (Utility and Limitations)

	Desirable Properties	$\mathrm{Uni}(Z:\hat{Y} \mid X_c)$	$\mathrm{I}(Z;\hat{Y} \mid X_c)$	$\mathrm{I}(Z;\hat{Y} \mid X_c, X')$
1.	Complete exemption if $X_c = X$.	Yes	Yes	Yes
2.	Detects unique information about Z in $\hat{Y}$ not in X_c.	Yes	Yes	Not Always
3.	Detects Non-Exempt Masked Discrimination.	No	Masked by $g(X_c)$	Masked by $g(X_c, X')$
4.	No causal influence from Z to $\hat{Y} \Rightarrow M_{NE} = 0$.	Yes	Not Always	Not Always

fairness (unlike equalized odds). The use of true labels has been criticized in Barocas and Selbst (2016) because "often the best labels for different classifications will be open to debate," e.g., if the labels themselves are biased. This work also shares intellectual connections with *individual fairness* (Dwork et al. 2012) in the sense that it enables individuals with similar X_c to be treated similarly, if desired.

Background on Partial Information Decomposition (PID): Here, we provide a brief background on the PID framework (Bertschinger et al. 2014; Williams and Beer 2010) to help follow this paper. The extended version (Dutta et al. 2019) provides more details and specific properties used in the proofs.

The PID framework decomposes the mutual information $\mathrm{I}(Z;(A,B))$ about a random variable Z contained in the tuple (A,B) into four *non-negative* terms as follows:

$$\mathrm{I}(Z;(A,B)) = \mathrm{Uni}(Z:A\backslash B) + \mathrm{Uni}(Z:B\backslash A) + \mathrm{Red}(Z:(A,B)) + \mathrm{Syn}(Z:(A,B)). \quad (1)$$

Here, $\mathrm{Uni}(Z:A\backslash B)$ denotes the unique information about Z that is present only in A and not in B. Likewise, $\mathrm{Uni}(Z:B\backslash A)$ is the unique information about Z that is present only in B and not in A. $\mathrm{Red}(Z:(A,B))$ denotes the redundant information about Z, present in both A and B, and $\mathrm{Syn}(Z:(A,B))$ denotes the synergistic information not present in either of A or B individually, but present jointly in (A,B) (see Dutta et al. (2019) for illustrations).

Example 1 (Partial Information Decomposition)**.** *Let $Z = (Z_1, Z_2, Z_3)$, $Z_i \sim$ i.i.d. Bern($1/2$). Let $A = (Z_1, Z_2, Z_3 \oplus N)$ where $\oplus$ denotes XOR, $B = (Z_2, N)$, and $N \sim$ Bern($1/2$) is independent of Z. Here, $\mathrm{I}(Z;(A,B)) = 3$ bits.*

Observe that, the unique information about Z that is contained only in A and not in B is effectively contained in Z_1 and is given by $\mathrm{Uni}(Z:A\backslash B) = \mathrm{I}(Z;Z_1) = 1$ bit. The redundant information about Z that is contained in both A and B is effectively contained in Z_2 and is given by $\mathrm{Red}(Z:(A,B)) = \mathrm{I}(Z;Z_2) = 1$ bit. Lastly, the synergistic information about Z that is not contained in either A or B alone, but is contained in both of them together is effectively contained in the tuple $(Z_3 \oplus N, N)$, and is given by $\mathrm{Syn}(Z:(A,B)) = \mathrm{I}(Z;(Z_3 \oplus N, N)) = 1$ bit. This accounts for the 3 bits in $\mathrm{I}(Z;(A,B))$. Here, B does not have any unique information about Z that is not contained in A.

Irrespective of the formal definition of these individual terms, the following identities also hold:

$$\mathrm{I}(Z;A) = \mathrm{Uni}(Z:A\backslash B) + \mathrm{Red}(Z:(A,B)). \quad (2)$$

$$\mathrm{I}(Z;A \mid B) = \mathrm{Uni}(Z:A\backslash B) + \mathrm{Syn}(Z:(A,B)). \quad (3)$$

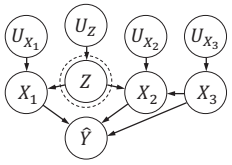

Figure 1: An SCM with protected attribute Z, features $X = \{X_1, X_2, X_3\}$, and output $\hat{Y}$. Z does not have any parents and $\hat{Y}$ is completely determined by $\{X_1, X_2, X_3\}$.

Given the three independent equations (1), (2) and (3) in four unknowns (the four PID terms), defining one of the terms (e.g., $\mathrm{Uni}(Z:A\backslash B)$) is sufficient to obtain the other three. For completeness, we include the definition of unique information from Bertschinger et al. (2014) (that also allows for estimation via convex optimization (Banerjee, Rauh, and Montúfar 2018)) with details in the extended version (Dutta et al. 2019). To follow the paper, only an intuitive understanding is sufficient.

Definition 1 (Unique Information)**.** *(Bertschinger et al. 2014) Let Δ be the set of all joint distributions on (Z,A,B) and Δ_p be the set of joint distributions with the same marginals on (Z,A) and (Z,B) as their true distribution, i.e., $\Delta_p = \{Q \in \Delta : q(z,a) = \Pr(Z{=}z, A{=}a)$ and $q(z,b) = \Pr(Z{=}z, B{=}b)\}$ Then, $\mathrm{Uni}(Z:A\backslash B) = \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;A|B)$.*

2 System Model and Assumptions

Definition 2 (Structural Causal Model: $\mathrm{SCM}(U,V,\mathcal{F})$)**.** *A structural causal model $(U,V,\mathcal{F})$ consists of a set of latent (unobserved) and mutually independent variables U which are not caused by any variable in the set of observable variables V, and a collection of deterministic functions (structural assignments) $\mathcal{F} = \{f_1, f_2, \ldots\}$, one for each $V_i \in V$, such that: $V_i = f_i(V_{pa_i}, U_i)$. Here $V_{pa_i} \subseteq V\backslash V_i$ are the parents of V_i, and $U_i \subseteq U$. The structural assignment graph (SAG) of $\mathrm{SCM}(U,V,\mathcal{F})$ has one vertex for each V_i, and directed edges to V_i from each parent in V_{pa_i}, and is always a directed acyclic graph.*

For our problem, the latent variables U represent possibly unknown social factors. The observables V consist of the protected attributes Z, the features $X = \{X_c, X_g\}$ and the output $\hat{Y}$ (see Fig. 1). For simplicity, we assume ancestral closure of the protected attributes, *i.e.*, the parents of any $V_i \in Z$ also lie in Z and hence Z is not caused by any of the features in X ($V_i \in Z$ are source nodes in the SAG).

Therefore, $Z = f_z(U_Z)$ for $U_Z \subseteq U$. Any feature X_j in X is a function of its corresponding latent variable and its parents, which are again functions of their own latent variables and parents. Note that, X can also be written as $f(Z, U_X)$ for some deterministic $f(\cdot)$, where $f(\cdot)$ may be constant in some of its arguments, and $Z \perp\!\!\!\perp U_X$ (see Peters, Janzing, and Schölkopf, Proposition 6.3). This holds because the underlying graph is acyclic. A model takes $X = \{X_c, X_g\}$ as its input and produces an output $\hat{Y}$ which depends only on X. Therefore, $\hat{Y} = h(Z, U_X)$ for some function $h(\cdot)$.

For completeness, we define Counterfactual Causal Influence (CCI) inspired from Kusner et al. (2017), Russell et al. (2017), Breiman (2001), Datta, Sen, and Zick (2016), Koh and Liang (2017), Adler et al. (2018), Henelius et al. (2014).

Definition 3 (Counterfactual Causal Influence: $\mathrm{CCI}(Z \to \hat{Y})$)**.** *If $\hat{Y} = h(Z, U_X)$ for some deterministic function $h(\cdot)$ where U_X are latent variables that do not cause Z in the true SCM, and Z', Z are $i.i.d.$, then*

$$\mathrm{CCI}(Z \to \hat{Y}) = \mathbb{E}_{Z,Z',U_X}\left[|h(Z, U_X) - h(Z', U_X)|\right].$$

Remark 1. *Statistical independence does not imply absence of causal effects. E.g., $\hat{Y} = Z \oplus U_X$ where $Z, U_X \sim$ i.i.d. Bern(½). Here, $\hat{Y} \perp\!\!\!\perp Z$, but Z still has a causal effect on $\hat{Y}$. If we vary Z while fixing all other sources of randomness in $\hat{Y}$ as constants (i.e., fixing $U_X = u_x$), then $\hat{Y}$ also varies. This is in fact an example of masked discrimination, where $\mathrm{I}(Z; \hat{Y}) = 0$, but Z causally influences $\hat{Y}$.*

Next, we define a variable W as follows:

Definition 4 (Variable W)**.** *We define a variable $W = [h(Z, u_x^{(1)}), \ldots, h(Z, u_x^{(k)})]$, where $\{u_x^{(1)}, \ldots, u_x^{(k)}\}$ is the set of all values with $\Pr(U_x = u_x) > 0$.*

Here, W is a deterministic function of Z alone, consisting of all the functional forms that $\hat{Y} = h(Z, U_X)$ takes for all values u_x attainable by U_X.

Lemma 1 (Information-Theoretic Equivalent of CCI)**.** *Let $\hat{Y} = h(Z, U_X)$ for some deterministic function $h(\cdot)$. Then $\mathrm{CCI}(Z \to \hat{Y}) \neq 0$ if and only if $\mathrm{I}(Z; W) > 0$.*

Remark 2. *We also show that $\mathrm{CCI}(Z \to \hat{Y}) = 0$ (or, $\mathrm{I}(Z; W) = 0$) is equivalent to the counterfactual fairness criterion of Kusner et al. (2017) (proved in the extended version (Dutta et al. 2019)). Therefore, in this work, we will regard $\mathrm{I}(Z; W)$ as an **information-theoretic quantification of the total discrimination (exempt and non-exempt)**.*

3 Main Results

We formally state the desirable properties, intuitively stated in Section 1, and then introduce our proposed measure that satisfies all of them (Theorem 1 in Section 3.1). While the proof is presented in the extended version (Dutta et al. 2019), in Section 3.2 we present the main intuition behind our proposed measure through several examples, counterexamples and thought experiments, that also help us arrive at the desirable properties. Our proposed measure leads to a non-negative decomposition of the total discrimination $\mathrm{I}(Z; W)$ into four components, i.e., statistically visible and masked portions, each with exempt and non-exempt components (see Section 3.3). Lastly, in Section 3.4, we demonstrate how to modify our measure to account for other kinds of masked discrimination under different sociological contexts.

3.1 Desirable Properties and Proposed Measure

We introduce a set of desirable properties for any measure of non-exempt discrimination (M_{NE}). Firstly, we require the measure to be 0 if all the features are in the exempt set X_c:

Property 1 (Complete Exemption)**.** *M_{NE} should be 0 if all features are categorized into X_c, i.e., $X_c = X$ and $X_g = \phi$.*

Next, it is desirable that the measure be non-zero if $\hat{Y}$ has any unique information about Z that is not present in X_c because then that information is also attributed to X_g.

Property 2 (Non-Exempt Visible Discrimination)**.** *M_{NE} should be strictly greater than 0 if $\mathrm{Uni}(Z : \hat{Y} \backslash X_c) > 0$.*

However, as we discussed in Section 2, statistical masking can sometimes prevent the entire non-exempt discrimination component from exhibiting itself in $\mathrm{Uni}(Z : \hat{Y} \backslash X_c)$. As an extreme example, consider the following scenario.

Example 2. *Let $\hat{Y} = Z \oplus f(U_X)$ for some function $f(\cdot)$ on $X_c = U_X$ with Z and $f(U_X)$ being i.i.d. Bern(½). E.g., an ad for expensive housing is presented to white people $(Z = 1)$ with income above a threshold $(f(U_X) = 1)$, and also to black people $(Z = 0)$ with income below a threshold $(f(U_X) = 0)$ (while being largely irrelevant to the latter).*

Not all forms of masked effects are undesirable. An example is if the only available features are $X_g = (Z, U_X)$, where Z is the race and U_X is Bern(½), a random coin flip. Then, performing $\hat{Y} = Z \oplus U_{X_1}$ randomizes the race, and can be a preventive measure against discrimination even if $\mathrm{CCI}(Z \to \hat{Y}) > 0$. In the following property, we will assume that the discrimination (masked/unmasked) is exempt if the Markov chain $Z - X_c - \hat{Y}$ holds. This property only accounts for masking that is entirely due to X_c, e.g., $\hat{Y} = Z + f(X_c)$ for some function $f(\cdot)$ where $\mathrm{CCI}(Z \to f(X_c)) = 0$ and exempts other forms of masking (revisited in Remark 3).

Property 3 (Non-Exempt Masking)**.** *A measure M_{NE} may be non-zero even if $\mathrm{I}(Z; \hat{Y}) = 0$. However, M_{NE} should be 0 if $Z - X_c - \hat{Y}$ form a Markov chain.*

Remark 3. *In general, one might also choose to consider a subset of latent factors $\widetilde{U} \subseteq U_X$ such that any statistical masking arising due to these latent variables is also undesirable. Then, the Markov chain in Property 3 may be modified to $Z - X_c - (\hat{Y}, \widetilde{U})$, and the proposed measure can be modified accordingly, as also elaborated further in Section 3.4.*

Lastly, the measure should also not capture false positives, e.g., it should be 0 if such virtual proxies cancel each other causing the final model output to have no counterfactual causal influence of Z, leading to the following property.

Property 4 (Cancellation of Influence)**.** *M_{NE} should be 0 if $\mathrm{CCI}(Z \to \hat{Y}) = 0$ (or equivalently, $\mathrm{I}(Z; W) = 0$).*

Now, we introduce our proposed measure and then show that it satisfies all these desirable properties (see Theorem 1).

Definition 5 (Non-Exempt Discrimination)**.** *Our proposed measure of non-exempt discrimination is given by:*

$$M_{NE} = \text{Uni}(Z : W \backslash X_c) - \text{Uni}(Z : W \backslash \hat{Y}, X_c). \quad (4)$$

Remark 4. *The proposed measure is essentially the volume of the overlap between* $\text{I}(Z; W)$ *and* $\text{I}(Z; \hat{Y}|X_c)$*, that becomes* 0 *when either of them is* 0 *(see Fig. 2).*

Theorem 1 (Properties)**.** *Properties 1, 2, 3 and 4 are satisfied by* $M_{NE} = \text{Uni}(Z : W \backslash X_c) - \text{Uni}(Z : W \backslash \hat{Y}, X_c)$.

3.2 Main Intuition behind the Proposed Measure

We examine some candidate measures (M_{NE}) of non-exempt discrimination through examples and counterexamples, leading to our proposed measure.

Candidate Measure 1. $M_{NE} = \text{I}(Z; \hat{Y})$.

Counterexample 1. *Let* $X_c = Z + U_{X_1}$ *where* $Z \sim$ *Bern(½),* $U_{X_1} \sim \mathcal{N}(0, \sigma_1^2)$ *and* $X_g = U_{X_2}$ *where* $U_{X_2} \sim \mathcal{N}(0, \sigma_2^2)$ *and is independent of* U_{X_1}*. The decision of the model is* $\hat{Y} = \text{sgn}\,(X_c + X_g - 0.5) = \text{sgn}\,(Z + U_{X_1} + U_{X_2} - 0.5)$.

Here, $\text{I}(Z; \hat{Y})$ is non-zero and so is $\text{I}(Z; X_c)$. However, $\text{I}(Z; \hat{Y}|X_c) = 0$ (see the Markov chain $Z - X_c - \hat{Y}$). The information that $\hat{Y}$ contains about Z is redundant information also contained in X_c. Therefore, the discrimination here should be exempted because it arises entirely from X_c.

Candidate Measure 2. $M_{NE} = \text{I}(Z; \hat{Y} \mid X_c)$.

This measure resolves Counterexample 1. It also has some provision for selectively capturing the non-exempt component: it is 0 in Counterexample 1, consistent with the intuition that there is no non-exempt discrimination. However, the following example exposes some of its limitations.

Counterexample 2 (Cancellation of Influence)**.** *Let* $X_c = Z + U_X$ *and* $X_g = Z$ *where* Z *denotes the gender and* U_X *denotes the student's knowledge. The model's decision on a student's ability is* $\hat{Y} = X_c - X_g = U_X$.

The influences of Z along two different causal paths cancel each other in the final output, so that $\text{CCI}(Z \to \hat{Y}) = 0$ (and, $\text{I}(Z; W) = 0$). Thus, there is no discrimination in the outcome $\hat{Y}$ (this is true even if the features in X_c were not exempt; see Remark 2). However, the measure $M = \text{I}(Z; \hat{Y}|X_c)$ is positive for this example, leading to a false positive in detecting discrimination. These two examples serve as our motivation behind Properties 1 and 4. The next candidate resolves both these examples.

Candidate Measure 3. $M_{NE} = \text{Uni}(Z : \hat{Y} \backslash X_c)$.

This measure resolves Counterexample 1: $\hat{Y}$ and X_c have redundant information about Z, but there is no unique information about Z in $\hat{Y}$ that is not in X_c. Thus $\text{Uni}(Z : \hat{Y} \backslash X_c) = 0$, consistent with the conclusion that the discrimination in Counterexample 1 should be exempt. $\text{Uni}(Z : \hat{Y} \backslash X_c)$ is also 0 in Counterexample 2. In fact, $\text{Uni}(Z : \hat{Y} \backslash X_c)$ captures the non-exempt discrimination that is statistically visible in $\text{I}(Z; \hat{Y})$, leading to Property 2.

Counterexample 3 (Masked Discrimination)**.** *Refer to Example 2 in Section 3.1 where* $\hat{Y} = Z \oplus f(X_c)$.

Here $Z \perp\!\!\!\perp \hat{Y}$, *i.e.*, $\text{I}(Z; \hat{Y}) = 0$, making the model "appear to have no discrimination." However, when examined more deeply, the model racially discriminates against half of the population (high-income black people) for whom the ad is relevant. This is also demonstrated by the fact that $\text{CCI}(Z \to \hat{Y}) \neq 0$ and the Markov chain $Z - X_c - \hat{Y}$ does not hold. $\text{Uni}(Z : \hat{Y} \backslash X_c)$ fails to capture such *non-exempt masked discrimination*. In fact, this example motivates Property 3. $\text{Uni}(Z : \hat{Y} \backslash X_c)$ does not satisfy this property as it has to be zero whenever $\text{I}(Z; \hat{Y}) = 0$.

Inspired from $\text{CCI}(Z \to \hat{Y})$, another possible candidate for quantifying non-exempt discrimination is a causal, path-specific examination (see also Chiappa and Gillam (2018), Kusner et al. (2017), Kilbertus et al. (2017)) by varying Z only along the direct paths through X_g and comparing if it causes any difference in the decision.

Candidate Measure 4. *Let* $\hat{Y} = h(Z, U_X)$ *in the true causal model. Assume a new causal graph with a new source node* Z' *having an independent and identical distribution as* Z *where we replace all direct edges from* Z *to* X_g *with an edge from* Z' *to* X_g*. Let* $\tilde{h}(Z, Z', U_X)$ *be the model output in the new causal graph. A candidate measure is* $M_{NE} = \mathbb{E}_{Z,Z',U_X}\left[|h(Z, U_X) - \tilde{h}(Z, Z', U_X)|\right]$.

Counterexample 4 (Non-zero Unique Information)**.** *Suppose that* $X_c = Z \oplus U_{X_1}$ *and* $X_g = U_{X_1}$ *where* Z *and* U_{X_1} *are i.i.d. Bern(½). Let* $\hat{Y} = X_c \oplus X_g = Z$.

In this example, $\hat{Y}$ has unique information about Z that is not contained in X_c, implying non-exempt visible discrimination. However, a path-specific examination would conclude that the causal influence of Z is only propagating through X_c, and hence should be exempt. Following the PID literature, here $\hat{Y}$ receives synergistic information about Z from both X_c and X_g, that cannot be attributed to X_c alone ($\text{I}(Z; X_c) = 0$). From an operational perspective, $\hat{Y}$ and X_c together lead to a better estimate of Z than X_c alone which means X_g is definitely a contributor to the discrimination, and thus $M_{NE} > 0$. We therefore seek a measure under which such discrimination qualifies as non-exempt. Motivated by this example, we now consider another candidate measure that is derived from $\text{I}(Z; W)$.

Candidate Measure 5. $M_{NE} = \text{Uni}(Z : W \backslash X_c)$.

While this measure resolves all the examples so far, it may not always satisfy Property 1.

Counterexample 5. *Suppose that* $X = X_c = Z \oplus U_X$*, and* $\hat{Y} = X_c = Z \oplus U_X$.

In this scenario, this measure is not 0 even though the discrimination is completely exempt. This motivates our proposed measure $M_{NE} = \text{Uni}(Z : W \backslash X_c) - \text{Uni}(Z : W \backslash \hat{Y}, X_c)$, which accounts for, and effectively removes, such exempt components from $\text{Uni}(Z : W \backslash X_c)$, and finally satisfies all the desirable properties.

M_{NE} being non-zero actually implies that both $\text{I}(Z; W) > 0$ and $\text{I}(Z; \hat{Y} | X_c) > 0$ (overlapping volume). However, this is only a one-way implication. $\text{I}(Z; W)$ and $\text{I}(Z; \hat{Y} | X_c)$ both being non-zero does not necessarily capture non-exempt discrimination.

Example 3. *Let* $Z = (Z_1, Z_2)$, $X_c = (Z_1 \oplus U_{X_1}, Z_2)$, $X_g = (Z_1, U_{X_2})$ *and* $\hat{Y} = (U_{X_1}, Z_2 \oplus U_{X_2})$ *where* $Z_1, Z_2, U_{X_1}, U_{X_2}$ *are i.i.d. Bern*(1/2).

This example should be exempt because Z_2 already appears in X_c, and is hence exempt. Our proposed measure also suggests the same conclusion. However, both $\text{I}(Z; W)$ and $\text{I}(Z; \hat{Y} | X_c)$ are non-zero here.

3.3 Understanding the Overall Decomposition

This work enables an information-theoretic decomposition of the total discrimination $\text{I}(Z; W)$ into non-exempt and exempt components, namely, M_{NE} and $\text{I}(Z; W) - M_{NE}$ respectively. Alongside, $\text{I}(Z; W)$ can also be decomposed into statistically visible and masked components, namely, $\text{I}(Z; \hat{Y})$ and $\text{I}(Z; W) - \text{I}(Z; \hat{Y})$ respectively. Combining these two decompositions leads to an overall four-way decomposition of $\text{I}(Z; W)$ as shown in Theorem 2 (see Fig. 2).

Theorem 2 (Non-Negative Decomposition of Total Discrimination)**.** *The total discrimination can be decomposed into four non-negative components as follows:*

$$\text{I}(Z; W) = M_{V,NE} + M_{V,E} + M_{M,NE} + M_{M,E}. \quad (5)$$

Here $M_{V,NE} = \text{Uni}(Z : \hat{Y} \backslash X_c)$ *is the visible, non-exempt component and* $M_{V,E} = \text{Red}(Z : (\hat{Y}, X_c))$ *is the visible, exempt component. These two terms add to form* $\text{I}(Z; \hat{Y})$ *which is the total statistically visible discrimination. Likewise,* $M_{M,NE} = M_{NE} - M_{V,NE}$ *is the masked, non-exempt component, and* $M_{M,E} = \text{I}(Z; W) - \text{I}(Z; \hat{Y}) - M_{M,NE}$ *is the masked, exempt component.*

The proof is in the extended version (Dutta et al. 2019).

Lemma 2 (Masked Discrimination)**.** *The total masked discrimination* $\text{I}(Z; W) - \text{I}(Z; \hat{Y})$ *is equal to* $\text{Uni}(Z : W \backslash \hat{Y})$.

Lemma 3 (Masked Discrimination Implications)**.** *The following two statements are equivalent:*

- $\text{I}(Z; \hat{Y} \mid U_X) - \text{I}(Z; \hat{Y}) > 0$.
- $\exists$ *a random variable* G *of the form* $G = g(U_X)$ *such that* $\text{I}(Z; \hat{Y} | G) > \text{I}(Z; \hat{Y})$.

Either of these statements imply $\text{I}(Z; W) - \text{I}(Z; \hat{Y}) > 0$.

3.4 Modifying the Proposed Measure to Account for More Masked Effects

Different forms of statistical masking can have different implications under different sociological contexts, e.g., $\hat{Y} = Z \oplus U_X$ may be undesirable if U_X is the income (recall Example 2) but not necessarily unfair if U_X is the random flip of a coin. In our proposed measure, we only accounted for statistical masking effects caused by the critical features X_c. However, there may be scenarios where we might want to capture masking effects by other variables also, e.g., X_g. Let us understand this using the following example.

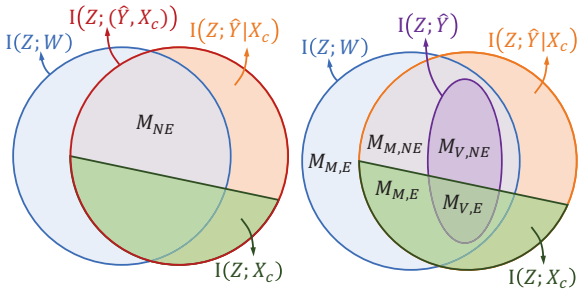

Figure 2: Information-theoretic decomposition of total discrimination, $\text{I}(Z; W)$: (Left) The red full-circle denotes $\text{I}(Z; (X_c, \hat{Y}))$ which is equal to $\text{I}(Z; X_c) + \text{I}(Z; \hat{Y} \mid X_c)$. Both $\text{I}(Z; X_c)$ and $\text{I}(Z; \hat{Y} \mid X_c)$ are denoted by subvolumes within the red full-circle. The volume of overlap between $\text{I}(Z; W)$ and $\text{I}(Z; \hat{Y} \mid X_c)$ is our proposed measure of non-exempt discrimination M_{NE}. (Right) Note that, $\text{I}(Z; \hat{Y})$ (total statistically visible discrimination) is is the purple circle that is entirely contained inside $\text{I}(Z; W)$ and $\text{I}(Z; \hat{Y} \mid X_c)$. This leads to a four-way decomposition of $\text{I}(Z; W)$: the visible non-exempt component $M_{V,NE} = \text{Uni}(Z : \hat{Y} \backslash X_c)$, the visible exempt component $M_{V,E} = \text{Red}(Z : (\hat{Y}, X_c))$, the masked non-exempt component $M_{M,NE} = M_{NE} - M_{V,NE}$, and the masked exempt component $M_{M,E} = \text{I}(Z; W) - \text{I}(Z; \hat{Y}) - M_{M,NE}$. Also note that $\text{I}(Z; \hat{Y})$ has an intersection with $\text{I}(Z; \hat{Y} | X_c)$, but both $\text{I}(Z; \hat{Y})$ and $\text{I}(Z; \hat{Y} | X_c)$ also have components (volumes) outside the intersection which allows either of them to be greater or less than the other in our Venn diagram.

Example 4. *Let* $X_c = (U_{X_1}, U_{X_2})$ *and* $X_g = (Z, U_{X_3})$, *where all the latent random variables are i.i.d. Bern*(1/2). *Now the output* $\hat{Y}$ *can take different forms, such as* $Z \oplus f_1(X_c) = Z \oplus U_{X_1}$, *or* $Z \oplus f_1(X_c) \oplus f_2(X_g) = Z \oplus U_{X_1} \oplus U_{X_3}$ *or* $Z \oplus f_2(X_g) = Z \oplus U_{X_3}$.

By our proposed measure, only $\hat{Y} = Z \oplus U_{X_1} \oplus U_{X_2}$ and $\hat{Y} = Z \oplus U_{X_1}$ are considered non-exempt. Masking by X_g (e.g., $\hat{Y} = Z \oplus U_{X_3}$) or masking by a combination of X_c and X_g (e.g., $\hat{Y} = Z \oplus U_{X_1} \oplus U_{X_3}$) is exempted ($Z - X_c - \hat{Y}$ is a Markov chain). Statistical masking of Z by $f_2(X_g)$ is viewed more like randomization, e.g., using a coin flip to prevent discrimination, whereas masking by $f_1(X_c)$ is like discriminating against high-income black people (Example 2).

In general, which masking effects should be accounted for depends on the problem design. In some scenarios, one may be interested in not exempting masking effects due to

some latent variables. Let $\widetilde{U}_X \subseteq U_X$ be the set of latent random variables such that any statistical masking effect derived from them should be accounted for. Then, we may redefine Property 3 as follows: the measure $M'_{NE} = 0$ if $Z - X_c - (\hat{Y}, \widetilde{U}_X)$ is a Markov chain. This leads to the following modified measure of non-exempt discrimination.

Definition 6 (Modified Non-Exempt Discrimination)**.** $M'_{NE} = \text{Uni}(Z : W \backslash X_c) - \text{Uni}(Z : W \backslash \hat{Y}, X_c, \widetilde{U}_X)$.

This measure is the volume of overlap between $\text{I}(Z; W)$ and $\text{I}(Z; \hat{Y}, \widetilde{U}_X \mid X_c)$. Using this measure in Example 4 leads to the conclusion that all the cases are non-exempt if $\widetilde{U}_X$ is chosen as $(U_{X_1}, U_{X_2}, U_{X_3})$. This unravels the statistical masking by $U_{X_1}, U_{X_2}, U_{X_3}$ and exposes the discriminatory component Z lying underneath. Again, in some examples, accounting for only some latent factors makes sense:

Example 5. *Let $X_c = Z + U_{X_1}$ and $X_g = (U_{X_1}, U_{X_2})$ where all the latent variables are independent with $U_{X_1} \sim \mathcal{N}(0, 1000)$ and all others distributed as $\mathcal{N}(0, 1)$. The output $\hat{Y}$ can take different forms, such as, $\hat{Y} = Z + U_{X_1}$, or $Z + U_{X_1} + U_{X_2}$, or $Z + U_{X_2}$.*

When $\hat{Y} = Z + U_{X_1}$, the output is entirely derived from X_c and hence should be exempt. Here, $Z - X_c - \hat{Y}$ is a Markov chain but $Z - X_c - (\hat{Y}, U_{X_1})$ is not. For this example, it does not make sense to try to unravel masked effects of U_{X_1} over Z, or include it in $\widetilde{U}_X$. When $\hat{Y} = Z + U_{X_1} + U_{X_2}$, it should also be exempt for the same reason. However, $\hat{Y} = Z + U_{X_2}$ is not necessarily exempt because it contains unique information about Z not present in X_c (X_g helps unmask and expose $Z + U_{X_2}$). Here, $Z - X_c - \hat{Y}$ is not a Markov chain. To unravel the masked effect caused by U_{X_2} and expose Z entirely, one may include it in $\widetilde{U}_X$.

4 Observational Relaxations for Practical Application in Training

Theorem 3 (Impossibility of Observational Measures)**.** *No observational measure of non-exempt discrimination simultaneously satisfies Properties 3 and 4.*

Nevertheless, because counterfactual measures are difficult to realize in practice, we examine the following observational measures of non-exempt discrimination that satisfy only a few of Properties 1-4.

1. $\text{Uni}(Z : \hat{Y} \backslash X_c)$: This measure satisfies Properties 1, 2 and 4 (proved in the extended version (Dutta et al. 2019)). However, it does not quantify any masked discrimination.

2. $\text{I}(Z; \hat{Y} | X_c)$: This measure satisfies Properties 1, 2, and 3 (proved in the extended version (Dutta et al. 2019)). However, it can lead to false positives for Property 4 (absence of $\text{CCI}(Z \to \hat{Y})$), e.g., in Counterexample 2.

3. $\text{I}(Z; \hat{Y} | X_c, X')$: X' consists of features of X_g suspected of masking Z. This is somewhat of a heuristic relaxation that only satisfies Property 1 but partly satisfies all the rest with some exceptions, *i.e.*, it exempts synergistic information about Z in (X_c, X') that can show up in $\hat{Y}$, and cause non-zero $\text{Uni}(Z : \hat{Y} \backslash X_c)$. It is able to detect more masked discrimination than $\text{I}(Z; \hat{Y} | X_c)$, i.e., when the mask is of the form $G = g(X_c, X')$. However, it can lead to false positives for Property 4 (absence of $\text{CCI}(Z \to \hat{Y})$).

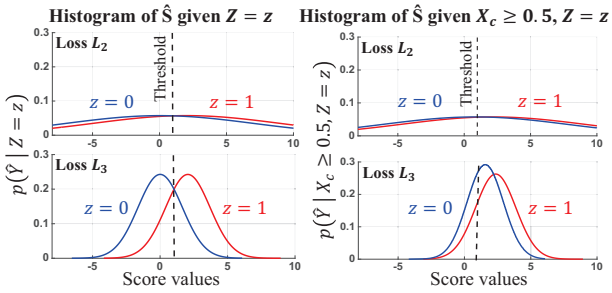

Figure 3: Histogram of Predicted Scores ($\hat{Y} = -w^T X/b$): (Left) $p(\hat{Y}|Z{=}i)$ for $i{=}0, 1$; (Right) $p(\hat{Y}|X_c{\geq}0.5, Z{=}i)$ for $i = 0, 1$. Regularizing with $\text{I}(Z; \hat{Y})$ (L_2) brings $p(\hat{Y}|Z)$ closer for $Z{=}0$ and 1 by placing higher weight on a less important feature (proximity score). This increases the variance and reduces accuracy (see Table 2). Regularizing with $\text{I}(Z; \hat{Y}|X_c)$ (L_3) makes $p(\hat{Y}|X_c{\geq}0.5, Z)$ approach each other for $Z{=}0$ and 1, aiming to give similar prediction scores to individuals with similar X_c ($\lambda{=}10$ for these plots).

Table 2: Observations after training a classifier ($w_1 X_1 + w_2 X_2 + w_3 X_3 + b \geq 0$) using three loss functions with different fairness criteria (100 simulations of 7000 iterations each with batch size 200).

Loss (λ)	$-\frac{w_1}{b}$	$-\frac{w_2}{b}$	$-\frac{w_3}{b}$	Acc%
L_1 $(-)$	1.08	1.08	1.08	98.5
L_2 (4)	1.07	1.07	3.76	81.1
L_2 (10)	1.01	1.03	13.9	70.2
L_3 (4)	1.46	0.73	1.91	89.6
L_3 (10)	2.05	0.02	2.57	80.8

Case Study: The goal is to decide whether to show ads for an editorial job requiring English proficiency, based on whether a score generated from internet activity is above a threshold. $Z \sim \text{Bern}(1/2)$ is a protected attribute denoting whether a person is a native English speaker or not. Now, consider three features $X = (X_1, X_2, X_3)$, such that: (i) X_1: a score based on online writing samples; (ii) X_2: a score based on browsing history, e.g., interest in English websites as compared to websites of other languages; and (iii) X_3: a preference score based on geographical proximity. Let $X_c = X_1$ and $X_g = (X_2, X_3)$.

Suppose the true SCM is as follows: $X_1 = Z + U_1$, $X_2 = Z + U_2$, $X_3 = U_3$ and the historic scores of selected candidates are $S = X_1 + X_2 + X_3$ where $U_1, U_2, U_3 \sim i.i.d.$ $\mathcal{N}(0, 1)$. Let the historic true labels be $Y = \mathbb{1}(S \geq 1)$ indicating whether $S \geq 1$ or not. We train a classifier of the form $\hat{Y} = 1/(1 + e^{-(w^T X + b)})$ (logistic regression). The classifier decides to show the ads if $\hat{Y} \geq 0.5$, *i.e.*, if $w^T X + b \geq 0$. We train using the following loss functions:

Loss L_1: $\min_{w,b} L_{\text{Cross Entropy}}(Y, \hat{Y})$.

Loss L_2: $\min_{w,b} L_{\text{Cross Entropy}}(Y,\hat{Y})+\lambda\widetilde{\mathrm{I}}(Z;\hat{Y})$, where λ is a regularizer and $\widetilde{\mathrm{I}}(Z;\hat{Y}) = -\frac{1}{2}\log(1-\rho^2_{Z,\hat{Y}})$ is an approximate expression of mutual information where $\rho_{Z,\hat{Y}}$ is the correlation between Z and $\hat{Y}$. This approximation is exact if Z and $\hat{Y}$ are jointly Gaussian.

Loss L_3: $\min_{w,b} L_{\text{Cross Entropy}}(Y,\hat{Y})+\lambda\widetilde{\mathrm{I}}(Z;\hat{Y}|X_c)$, where the range of X_c is first divided into multiple discrete bins, and $\widetilde{\mathrm{I}}(Z;\hat{Y}|X_c)$ is $\sum_i \Pr(X_c{\in}\text{Bin } i)\widetilde{\mathrm{I}}(Z;\hat{Y}|X_c{\in}\text{Bin } i)$ $=-\frac{1}{2}\sum_i \Pr(X_c{\in}\text{Bin } i)\log(1-\rho^2_{Z,\hat{Y},i})$ and $\rho_{Z,\hat{Y},i}$ is the conditional correlation of $\hat{Y}$ and Z given that X_c is in the i-th discrete bin.

Observations: For L_1, the separation boundary is very close to that based on the historic scores. But, because the past scores are correlated with browsing history (X_2), there is a danger that even when a non-native speaker has good writing score, they may not be shown an ad due to their browsing history. Regularizing with $\mathrm{I}(Z;\hat{Y})$ (Loss L_2) does not work well because the model begins to weigh both X_1 and X_2 less, and many proficient candidates are dropped in favour of a less-important feature, namely, proximity (X_3), also reducing the accuracy (see Table 2). However, regularizing with $\mathrm{I}(Z;\hat{Y} \mid X_c)$ (Loss L_3) is able to reduce the importance (weight) of browsing history relative to online scores, leading to an intermediate accuracy between L_1 and L_2 for same λ (see also Fig. 3). In a sense, our measure enables individuals with similar X_c to be treated similarly.

5 Discussion

This work provides a novel information-theoretic quantification of fairness under exemptions by adopting an axiomatic approach. We note that our properties, as stated, do not lead to a unique measure of non-exempt discrimination. They provide a qualitative separation of exempt and non-exempt discrimination, but, in line with much of the literature on fairness, do not quantify its "scaling." However, it is not obvious what properties one can use to constrain this scaling, and remains an open question to pursue as future work. In fact, we believe that there is value in the fact that the properties do not yield a unique measure: this allows for tuning the measure for the needs of an application. E.g., Shannon established uniqueness on Shannon's entropy with respect to some properties in Shannon (1948) but the needs of the application can still drive the use of alternate measures, e.g. Renyi entropy (Rényi 1961) that weighs outliers differently than Shannon entropy.

While our properties do not quantify the scaling, the measure we propose does capture important aspects of the problem, e.g., it captures both masked and statistically visible components when they are present together, that existing measures such as $\mathrm{I}(Z;\hat{Y})$ or $\mathrm{Uni}(Z : \hat{Y}\backslash X_c)$ do not. E.g., let $X_c = U \sim \mathcal{N}(0,1)$, $X_g = Z \sim \text{Bern}(1/2)$, and $\hat{Y} = Z+U$, i.e., Z is partially masked by U even though the visible discrimination is nonzero (a modification of Counterexample 3). Here, our measure is equal to the Shannon entropy $\mathrm{H}(Z)$, whereas $\mathrm{I}(Z;\hat{Y})$ or $\mathrm{Uni}(Z : \hat{Y}\backslash X_c)$ are smaller than $\mathrm{H}(Z)$ because they do not account for the masked component.

We also acknowledge that given the probability distribution on the data, an SCM is not always unique (Peters, Janzing, and Schölkopf 2017) making it difficult to use counterfactual measures in practice (as also noted for other results in the field, e.g., Kilbertus et al. (2017), Kusner et al. (2017)). To address this, we also propose observational relaxations of our measure and analyze what they capture and what they miss (see Table 1). In practice, this can inform which measure can be used when, e.g., $\mathrm{I}(Z;\hat{Y}|X_c)$ can be used when cancellation of influences (Counterexample 2) does not occur (i.e., if the SCM satisfies certain faithfulness assumptions). Similarly, $\mathrm{Uni}(Z : \hat{Y}\backslash X_c)$ may be used when accounting for masked discrimination is not required. Since the assumptions in relaxing the measure to observational ones are explicitly identified, corrections can be made if it is found that these assumptions are not satisfied. Finally, in scenarios where the SCM is known or can be evaluated from the data (see Chapters 4 and 7 in Peters, Janzing, and Schölkopf (2017)), the proposed measure exactly captures the non-exempt discrimination.

References

Adler, P.; Falk, C.; Friedler, S. A.; Nix, T.; Rybeck, G.; Scheidegger, C.; Smith, B.; and Venkatasubramanian, S. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54(1):95–122.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Anonymous. 2019. Conditional Debiasing for Neural Networks.

Banerjee, P. K.; Rauh, J.; and Montúfar, G. 2018. Computing the unique information. In *IEEE International Symposium on Information Theory*, 141–145.

Barocas, S., and Selbst, A. D. 2016. Big data's disparate impact. *Calif. L. Rev.* 104:671.

Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; and Ay, N. 2014. Quantifying unique information. *Entropy* 16(4):2161–2183.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, 3992–4001.

Chiappa, S., and Gillam, T. P. 2018. Path-specific counterfactual fairness. *arXiv preprint arXiv:1802.08139*.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, 797–806. ACM.

Datta, A.; Fredrikson, M.; Ko, G.; Mardziel, P.; and Sen, S. 2017. Use privacy in data-driven systems: Theory and ex-

periments with machine learnt programs. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1193–1210.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy*, 598–617.

Dept. of Trade and Industry. 2008. Genuine Occupational Qualifications, A Good Practice Guide for Employers.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J. S.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2791–2801.

Dutta, S.; Venkatesh, P.; Mardziel, P.; Datta, A.; and Grover, P. 2019. An information-theoretic quantification of discrimination with exempt features. https://sites.google.com/site/sanghamitraweb/academic-articles.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Ghassami, A.; Khodadadian, S.; and Kiyavash, N. 2018. Fairness in supervised learning: An information theoretic approach. In *IEEE International Symposium on Information Theory*, 176–180.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323.

Heidari, H.; Ferrari, C.; Gummadi, K.; and Krause, A. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, 1265–1276.

Henelius, A.; Puolamäki, K.; Boström, H.; Asker, L.; and Papapetrou, P. 2014. A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery* 28(5-6):1503–1529.

Kamiran, F.; Žliobaitė, I.; and Calders, T. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems* 35(3):613–644.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50. Springer.

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.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 1885–1894. JMLR. org.

Komiyama, J., and Shimao, H. 2017. Two-stage algorithm for fairness-aware machine learning. *arXiv preprint arXiv:1710.04924*.

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

Liao, J.; Huang, C.; Kairouz, P.; and Sankar, L. 2019. Learning generative adversarial representations (gap) under fairness and censoring constraints. *arXiv preprint arXiv:1910.00411*.

Manley, K. 2009. The bfoq defense: Title vii's concession to gender discrimination. *Duke J. Gender L. & Pol'y* 16:169.

Menon, A. K., and Williamson, R. C. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, 107–118.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. MIT press.

Rényi, A. 1961. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1. The Regents of the University of California.

Russell, C.; Kusner, M. J.; Loftus, J.; and Silva, R. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, 6414–6423.

Salimi, B.; Rodriguez, L.; Howe, B.; and Suciu, D. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, 793–810. ACM.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell system technical journal* 27(3):379–423.

Tax, T.; Mediano, P.; and Shanahan, M. 2017. The partial information decomposition of generative neural network models. *Entropy* 19(9):474.

US Equal Employment Opportunity Commission. 1963. The Equal Pay Act of 1963. https://www.eeoc.gov/laws/statutes/epa.cfm.

Varshney, K. R. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25(3):26–29.

Venkatesh, P.; Dutta, S.; and Grover, P. 2019. Information flow in computational systems. *arXiv preprint arXiv:1902.02292*.

Wang, H.; Ustun, B.; and Calmon, F. P. 2019. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. *arXiv preprint arXiv:1901.10501*.

Williams, P. L., and Beer, R. D. 2010. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, 325–333. JMLR. org.