Fairness Under Feature Exemptions: Counterfactual and Observational Measures

Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, Pulkit Grover

Abstract

With the growing use of machine learning algorithms in highly consequential domains, the quantification and removal of bias with respect to protected attributes, such as gender, race, etc., is becoming increasingly important. While quantifying bias is essential, sometimes the needs of a business (e.g., hiring) may require the use of certain features that are critical in a way that any bias that can be explained by them might need to be exempted. For instance, in hiring an employee, a standardized test-score may be a critical feature that should be weighed strongly in the decision even if biased, whereas other features, such as name, zip code, or reference letters may be used to improve decision-making, but only to the extent that they do not reinforce bias. In this work, we propose a novel information-theoretic decomposition of the total bias (a quantification inspired from counterfactual fairness) into a non-exempt component which quantifies the part of the bias that cannot be accounted for by the critical features, and an exempt component which quantifies the remaining bias. This decomposition is important: it allows one to check if the bias arose purely due to the critical features (inspired from the business necessity defense of disparate impact law) and also enables selective removal of the non-exempt component of bias if desired. We arrive at this decomposition through examples and counterexamples that lead to a set of desirable properties (axioms) that any measure of non-exempt bias should satisfy. We then demonstrate that our proposed counterfactual measure of non-exempt bias satisfies all of them. Our quantification bridges ideas of causality, Simpson's paradox, and 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 bias 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 case studies to show how one can train models while reducing non-exempt bias.

I. Introduction

As artificial intelligence becomes ubiquitous, it is important to understand whether a machine-learnt model is unfairly biased with respect to *protected attributes* such as gender, race, etc., and if so, how we can engineer fairness into such a model. The field of fair machine learning provides several measures for fairness [2]–[40], and uses them to reduce bias, e.g., as a regularizer during training [3], [7]. In several 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. Examples of such critical features include weightlifting ability for a firefighter's job, educational qualification for an academic job, coding skills for a software engineering job, merit and seniority in deciding salary, etc. In an attempt to preserve the importance of the critical features in the decision making, one might choose to exempt the bias arising from them. On the other hand, racial bias in mortgage lending decisions arising from zip code (a non-critical feature) [41] or gender bias in automated hiring arising from the word "women's" in a resume [42] are examples of non-exempt bias. In this work, our goal is to formalize and quantify the *non-exempt bias*, i.e., the part of the bias that cannot be accounted for by the critical features. This quantification is important for two main reasons: (i) it allows one to check if the bias arose purely due to the critical features (inspired from the "business necessity defense" in the disparate impact law, i.e., Title VII of the Civil Rights Act of 1964 [43]); and (ii) it enables selective removal of the non-exempt component if desired.

In this work, we assume that the critical features or business necessities are known (similar to [19], [44]; this discussion is revisited in Section VIII). We let X_c and X_g denote the critical and the non-critical (or general) features, and X denote the entire set of features. We also denote the protected attribute(s) by Z, the true label by Y, and the model output by $\hat{Y}$ which is a function of the entire feature vector X. While we acknowledge that such categorization of features is application-dependent and might require domain knowledge and ethical evaluation, such exemptions do exist in law. E.g., the US Equal Pay Act [45] exempts for difference in salary based on gender that can be explained by merit and seniority. Similarly, the US employment discrimination law [46] contains a Bona Fide Occupational Qualification (BFOQ) defense where bias about protected attributes may be exempted if the bias is "due to a BFOQ reasonably necessary to the normal operation of that particular business or other reasonable differentials." For example, weightlifting ability is a critical feature in hiring firefighters so that they are able to carry fire victims out of a burning building. The feature representing weightlifting ability is therefore required to be weighed strongly in hiring even if it is correlated with some protected attributes. Similarly, UK employment bias 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 [47].

Why should we 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% applicants clear a test, but resources are available to interview only 10%. Not using the general features at all can reduce accuracy, or produce a very large candidate pool. In this work, our proposition is

The authors are with the Department of Electrical and Computer Engineering, Carnegie Mellon University. Author Contacts: S. Dutta (sanghamd@andrew.cmu.edu), P. Venkatesh (vpraveen@cmu.edu), P. Mardziel (piotrm@cmu.edu), A. Datta (danupam@cmu.edu), P. Grover (pulkit@cmu.edu).
This work has been presented in part at AAAI 2020 [1].

TABLE I: Observational Measures (M_{NE}) of Non-Exempt Bias (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.	M_{NE} is equal to total bias (in a counterfactual sense) if $X_c = \phi$ and $X_g = X$.	No	No	No
2.	M_{NE} is non-increasing as more features are added to X_c from X_g.	Yes	No	No
3.	M_{NE} is 0 (complete exemption) if $X_c = X$ and $X_g = \phi$.	Yes	Yes	Yes
4.	No counterfactual causal influence from Z to $\hat{Y} \Rightarrow M_{NE} = 0$.	Yes	Not Always	Not Always
5.	M_{NE} detects unique information about Z in $\hat{Y}$ not in X_c.	Yes	Yes	Not Always
6.	M_{NE} detects non-exempt masked bias.	No	Masked by $g(X_c)$	Masked by $g(X_c, X')$

to use both critical and general features in a way that maximizes accuracy (to the extent possible) while preventing non-exempt bias. For instance (inspired from [43]), to choose a "good" employee, an employer could evaluate standardized test scores and also reference letters (human-graded performance reviews). All these features are "job-related" in that they have statistical correlation with the prediction goal, and can help improve the 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.

This work treads a middle ground between two popular measures of fairness that do not use domain knowledge, namely, *statistical parity* [2], [3], [10], [35], which enforces the criterion $Z \perp\!\!\!\perp \hat{Y}$, and *equalized odds* [4], [10], [35], which enforces $Z \perp\!\!\!\perp \hat{Y}|Y$ (directly or through practical relaxations). Our selective quantification of non-exempt bias (using domain knowledge to identify critical features) helps address one of the major criticisms against statistical parity. The criticism is that it can lead to the selection of unqualified members from the protected group [4], [26], e.g., by disregarding the critical features if they are correlated with the protected attribute Z. In fact, in our case study in Section VII, we observe that the weight of the critical feature is significantly reduced in the decision making when one uses statistical parity as a regularizer with the loss function because the critical feature is correlated with Z. On the other hand, equalized odds suffers from label bias [36], [37], [41], [48], [49] because it is based on agreement with the true labels. In fact, we will demonstrate through an example (Example 5 in Section III-C) that if the historic labels themselves reinforce bias from the non-critical features, then even if we obtain a perfect classifier after training on the historic data, which satisfies equalized odds, it can reinforce undesirable non-exempt bias[1].

A. Contributions

Our main contribution in this work is the quantification of non-exempt bias based on a rigorous axiomatic approach. As a first step towards this quantification, we propose an information-theoretic quantification (see Definition 4 in Section II-B) of the total bias (exempt and non-exempt) that is 0 if and only if the model is *counterfactually fair* [18]. Counterfactual fairness [18], [20] is a causal notion of fairness where the features X, the protected attribute Z and the model output $\hat{Y}$ are assumed to be observables in a Structural Causal Model (SCM) (defined formally in Section II; see Definition 2). The model is deemed *counterfactually fair* if Z has no *counterfactual causal influence* on $\hat{Y}$, i.e., $\hat{Y}$ does not change if we are able to vary Z in the SCM in a manner that other independent latent factors remain constant (defined formally in Section II; see Definition 3).

Interestingly, note that the total bias (in a counterfactual sense) may not exhibit itself entirely in the mutual information $\mathrm{I}(Z; \hat{Y})$, which is the *statistically visible information*[2] about Z in $\hat{Y}$, because of "statistical masking effects" (also relates to Simpson's paradox [50]). Consider an example inspired from [18], [22], [34] where an expensive housing ad is shown selectively to high-income people of one race and also to low income people of another race, i.e., $\hat{Y} = Z \oplus G$ where $\oplus$ denotes XOR, G is the income that has no causal influence of Z and G, Z are i.i.d. Bern(1/2). This model is biased against the high-income people of the second race for whom the ad is relevant, but $\mathrm{I}(Z; \hat{Y}) = 0$ here, thus failing to capture this bias. Intuitively, our quantification of total bias also extends the idea of *proxy-use* [22] from *white-box models*[3] to black-box models. Proxy-use [22] examines "white-box" models, i.e., models with clearly defined constituents (e.g., decision trees) and regards a model as having bias if (i) there is a constituent that has high mutual information about Z (a proxy of Z); and (ii) this constituent also causally influences the output $\hat{Y}$ (i.e., varying the constituent while keeping other constituents constant does not change the output). In this work, the total bias captures the intuitive notion of a virtual constituent or proxy of Z that causally influences the final output $\hat{Y}$ (this intuition is revisited to understand Example 2 in Section II-B). For instance, a virtual constituent Z is formed in the example of masked bias in housing ads that causally influences $\hat{Y}$ even though $\mathrm{I}(Z; \hat{Y}) = 0$.

Next, we quantify the *non-exempt* part of this total bias, i.e., the part that cannot be explained by the critical features (X_c). Building on the extension of proxy-use [22] for black-box models as discussed above, we aim to quantify the influence of a discriminatory virtual constituent or proxy of Z, if formed inside the black-box model, on the model output $\hat{Y}$, and that

[1]Our quantification does not use the true labels for fairness (unlike equalized odds), addressing the criticism in [43] which says that " [...] often the best labels for different classifications will be open to debate."

[2]This is a quantification of bias inspired from statistical parity which deems a model fair if and only if $\hat{Y} \perp\!\!\!\perp Z$. Note that, $\mathrm{I}(Z; \hat{Y}) = 0$ if and only if $\hat{Y} \perp\!\!\!\perp Z$.

[3]White-box models [22] are the type of models where one can clearly explain how they behave, how they produce predictions and what the influencing variables or sub-components of the model are, e.g., decision trees, linear regression, etc.

cannot be attributed entirely to the critical features (this idea is revisited for an intuitive understanding of the examples in Section II-B.). To quantify this *non-exempt bias*, we consider toy examples and thought experiments to first arrive at a set of desirable properties (axioms) that any measure of non-exempt bias should satisfy, and then provide a measure that satisfies them (see Theorem 1). These desirable properties can be intuitively described as follows. First, if all the features are in the non-critical set, then the measure should be equal to the total bias since no bias is exempt. For a fixed set of features X and a fixed model, as more features become categorized as critical, the measure of non-exempt bias should not increase, i.e., it either decreases or stays the same. Ultimately, if all the features are in the critical set X_c, then we require the measure of non-exempt bias to be 0 since then the total bias is exempt. The measure should also not avoid false positive conclusions, e.g., it should be 0 if the virtual constituents or proxies of Z cancel each other leading to a final model output that has no counterfactual causal influence of Z, i.e., the model is counterfactually fair. Next, it is desirable that the measure be non-zero if $\hat{Y}$ has any "unique" statistically visible 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 bias* that needs to be captured, e.g., in the aforementioned example of housing ads (also revisited in Example 3 in Section III-A where we formally state these properties).

Our proposed measure of non-exempt bias, that satisfies all these desirable properties, is *counterfactual* in nature, i.e., it depends on the true SCM, and hence, is not *observational*[4] in general. We also show the theoretical impossibility of any observational measure in satisfying all the desirable properties together (see Theorem 3). We note that in some applications, counterfactual measures can be realized or approximated with assumptions on the causal model. However, for more general use in practical applications, we also propose several observational relaxations of our measure that satisfy only some of these properties. Nevertheless, we believe that a counterfactual measure and its properties are crucial in understanding the utility and the limitations of different observational measures and informing which measure to choose in practice (summarized in Table I; detailed discussion in Section VI).

To summarize, our contributions in this work are as follows:

1. Quantification of Non-Exempt Bias: We propose a novel counterfactual measure of non-exempt bias that captures the bias that cannot be explained by the critical features. Our quantification attempts to capture the intuitive notion of whether a discriminatory virtual constituent or proxy [22] of Z is formed inside the black-box model that influences the output $\hat{Y}$ and that cannot be attributed entirely to the critical features (X_c). We adopt a rigorous axiomatic approach where we first arrive at a set of desirable properties that any measure of non-exempt bias should satisfy, and then show that the proposed measure satisfies them (see Theorem 1). Our quantification leverages a body of work in information theory called Partial Information Decomposition (PID), as well as, works on counterfactual fairness.

2. Overall Decomposition of Total Bias into Statistically Visible and Masked components: Our quantification finally leads us to an overall decomposition of the total bias into four non-negative components, namely, exempt and non-exempt *statistically visible* bias and exempt and non-exempt *masked* bias (see Theorem 2). The exempt and non-exempt *statistically visible* biases add up to give $\mathrm{I}(Z;\hat{Y})$ which is the total statistically visible bias.

3. An Impossibility Result: We show that no purely observational measure of non-exempt bias 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 I) and then use one of them, namely, conditional mutual information, to demonstrate how to selectively reduce non-exempt bias in practice through case studies.

Related Works: Causal approaches for fairness have been explored in [18]–[22], including impossibility results on purely observational measures [19], [22]. Our main novelty lies in our adoption of a rigorous axiomatic approach based on examples and thought experiments for quantification of non-exempt bias while allowing for exemptions due to critical features. Our quantification of non-exempt bias attempts to capture the intuitive notion of virtual constituents or proxies of Z forming inside the black box model that cannot be attributed to the critical features X_c alone (inspired from proxy-use [22]). The decomposition of total bias into exempt and non-exempt components is tricky. For instance, following the ideas of path-specific counterfactual fairness [21], 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, examples from the PID literature show that bias can also arise from synergistic information 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 be (see Counterexample 3). Purely causal measures (that do not rely on the PID framework) can attribute such bias entirely to X_c. We contend that such synergistic information, if influencing the decision, must be included in the *non-exempt* component of bias because both X_c and X_g are contributors. We note that identifying such synergy is important: synergy arises frequently in machine-learning and other related applications [50]–[54].

We are also aware that the idea of using mutual information or dependence conditioned on the critical feature(s) as a measure of non-exempt bias has surfaced in another work [55], where the focus is on conditional debiasing of neural networks using novel estimators. Other observational measures of non-exempt bias that examine the dependence between Z and $\hat{Y}$ after

[4]Observational measures are those that can be estimated from the probability distribution of the data without knowledge of the underlying SCM.

conditioning on subsets of features or other sub-groups have been used in [34], [44], [56], [57]. In this work, our focus is on theoretical understanding of such observational measures (existing measures as well as observational relaxations of our counterfactual measure) and relating them information-theoretically with *counterfactual fairness*[5] [18], [20], which has not received detailed attention. Indeed, we provide example decisions (e.g., see Counterexample 1) which may be deemed *unfair* by observational measures based on conditioning, e.g., conditional mutual information (or conditional statistical parity), but are deemed *fair* by counterfactual fairness [18], [20]. In light of our examples, we examine some related observational measures, namely, justifiable fairness [56], and conditional statistical parity [55], [57] as well as the related causal measure of path-specific counterfactual fairness [21] in Section III-C to understand what they capture and what they miss. We also discuss the utility and the limitations of different observational relaxations of our proposed counterfactual measure in Section VI (e.g., see an impossibility result on observational measures in Theorem 3) followed by some case studies in Section VII and a concluding discussion in Section VIII.

B. Paper Outline

The rest of the paper is organized as follows. Section II introduces the background, system model and assumptions underlying our problem formulation, i.e., how to quantify the non-exempt bias. Section III-A first states all the desirable properties that a measure of non-exempt bias should satisfy, and then introduces our proposed counterfactual measure that satisfies all of them (Theorem 1 in Section III-A). This is followed by a rationale behind the desirable properties through examples and thought experiments in Section III-B. Our examples also demonstrate the utility and limitations of some existing measures, namely, conditional statistical parity [57], path-specific counterfactual fairness [21], and justifiable fairness [56], as we discuss in Section III-C. Next, Section IV provides insights on the overall decomposition of the total bias (in a counterfactual sense) into exempt and non-exempt components, with each of them being further decomposed into *statistically visible* and *masked* components (Theorem 2 in Section IV). Section V provides an impossibility result on observational measures, stating that no observational measure can satisfy all of the desirable properties. Nonetheless, since counterfactual measures are often difficult to realize in practice, we propose several observational relaxations of our proposed counterfactual measure in Section VI (that only satisfy some of the desirable properties), and discuss their utility and limitations. Next, in Section VII, we use one of our observational measure to conduct case studies on both artificial and real datasets to demonstrate practical application in training. Finally, we conclude with a discussion in Section VIII.

II. Preliminaries

Here, we first provide a brief background on Partial Information Decomposition (PID) in Section II-A to help follow the paper. Appendix B provides more details on the specific properties used in the proofs. Next, we introduce our system model and assumptions in Section II-B. We use the following notations: (i) $X=(X_1,X_2,\dots,X_n)$ denotes a tuple [58], i.e., an ordered set of elements $X_1,X_2,\dots,X_n$; (ii) ϕ denotes the empty tuple (no elements); (iii) For tuple with a single element, the bracket is omitted for brevity, i.e., $(X_1)=X_1$; (iv) (X,A) is equivalent to the new tuple $(X_1,X_2,\dots,X_n,A)$ formed by appending the element A at the end of tuple X; (v) $X_1\in X$ means X_1 is an element of tuple X; (vi) $S\subseteq X$ means the set of elements in tuple S form a subset of the set of elements in tuple X; and (vii) $X\backslash X_2$ denotes a new tuple formed by removing element X_2 from X without changing the order of other elements, i.e., $(X_1,X_3,X_4,\dots,X_n)$.

A. Background on Partial Information Decomposition (PID)

The PID framework [59]–[61] 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 (also see Fig. 1):

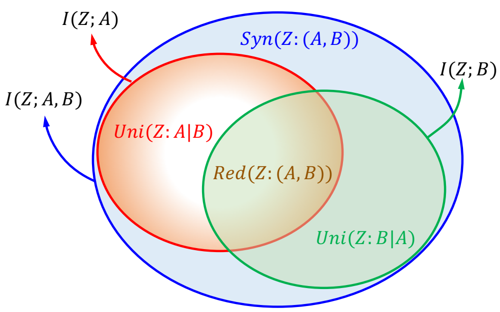
$$\mathrm{I}(Z;(A,B))=\mathrm{Uni}(Z:A|B)+\mathrm{Uni}(Z:B|A)+\mathrm{Red}(Z:(A,B))+\mathrm{Syn}(Z:(A,B)). \tag{1}$$

Here, $\mathrm{Uni}(Z:A|B)$ denotes the unique information about Z that is present only in A and not in B. Likewise, $\mathrm{Uni}(Z:B|A)$ is the unique information about Z that is present only in B and not in A. The term $\mathrm{Red}(Z:(A,B))$ denotes the redundant information about Z that is 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). *All four of these terms are non-negative. Also notice that,* $\mathrm{Red}(Z:(A,B))$ *and* $\mathrm{Syn}(Z:(A,B))$ *are symmetric in* A *and* B. Before defining these PID terms formally, let us understand an intuitive example.

Example 1 (Partial Information Decomposition)**.** *Let $Z=(Z_1,Z_2,Z_3)$ with $Z_1,Z_2,Z_3\sim$ i.i.d. Bern(½). Let $A=(Z_1,Z_2,Z_3\oplus N)$, $B=(Z_2,N)$, $N\sim$ Bern(½) is independent of Z. Here, $\mathrm{I}(Z;(A,B))=3$ bits.*

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 $\mathrm{Uni}(Z:A|B)=\mathrm{I}(Z;Z_1)=1$ 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

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



(a) Venn diagram showing PID of $\mathrm{I}(Z;(A,B))$

	$I(Z;A\mid B)$	$I(Z;B)$	$I(Z;A,B)$
$I(Z;A)$	$Uni(Z:A\mid B)$	$Red(Z:(A,B))$	
$I(Z;B\mid A)$	$Syn(Z:(A,B))$	$Uni(Z:B\mid A)$	
$I(Z;A,B)$			

(b) Tabular Representation of PID of $\mathrm{I}(Z;(A,B))$

Fig. 1: Mutual information $\mathrm{I}(Z;(A,B))$ is decomposed into 4 non-negative terms, namely, $\mathrm{Uni}(Z:A|B)$, $\mathrm{Uni}(Z:B|A)$, $\mathrm{Red}(Z:(A,B))$ and $\mathrm{Syn}(Z:(A,B))$. Also note that, $\mathrm{I}(Z;(A,B)) = \mathrm{I}(Z;B) + \mathrm{I}(Z;A \mid B)$, each of which is in turn a sum of two PID terms. $\mathrm{Red}(Z:(A,B))$ is the sub-volume between $\mathrm{I}(Z;A)$ and $\mathrm{I}(Z;B)$, and $\mathrm{Uni}(Z:A|B)$ is the sub-volume between $\mathrm{I}(Z;A \mid B)$ and $\mathrm{I}(Z;A)$.

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, i.e., $\mathrm{Uni}(Z:B|A) = 0$.

Irrespective of the formal definition of these individual terms, the following identities also hold (see Fig. 1b):

$$\mathrm{I}(Z;A) = \mathrm{Uni}(Z:A|B) + \mathrm{Red}(Z:(A,B)). \tag{2}$$
$$\mathrm{I}(Z;A \mid B) = \mathrm{Uni}(Z:A|B) + \mathrm{Syn}(Z:(A,B)). \tag{3}$$

Remark 1 (An Interpretation of PID as Information-Theoretic Sub-volumes)**.** *Equations* (1), (2) *and* (3) *have been represented in a tabular fashion in Fig. 1b. Notice that,* $\mathrm{Uni}(Z:A|B)$ *can be viewed as the information-theoretic sub-volume of the intersection between* $\mathrm{I}(Z;A)$ *and* $\mathrm{I}(Z;A \mid B)$*. Similarly,* $\mathrm{Red}(Z:(A,B))$ *is the sub-volume between* $\mathrm{I}(Z;A)$ *and* $\mathrm{I}(Z;B)$*.*

These equations also demonstrate that $\mathrm{Uni}(Z:A|B)$ and $\mathrm{Red}(Z:(A,B))$ are the information contents that exhibit themselves in $\mathrm{I}(Z;A)$ which is the statistically visible information content about Z present in A. Because both these PID terms are non-negative, if any one of them is non-zero, we will have $\mathrm{I}(Z;A) > 0$. Similarly, $\mathrm{Uni}(Z:B|A)$ and $\mathrm{Red}(Z:(A,B))$ also exhibit themselves in $\mathrm{I}(Z;B)$. On the other hand, $\mathrm{Syn}(Z:(A,B))$ is the information content that does not exhibit itself in $\mathrm{I}(Z;A)$ or $\mathrm{I}(Z;B)$ individually, i.e., these terms can still be 0 even if $\mathrm{Syn}(Z:(A,B)) > 0$. But, $\mathrm{Syn}(Z:(A,B))$ exhibits itself in $\mathrm{I}(Z;(A,B))$. Notice that,

$$\begin{aligned}\mathrm{I}(Z;(A,B)) &= \underbrace{\mathrm{Uni}(Z:A|B) + \mathrm{Red}(Z:(A,B))}_{\mathrm{I}(Z;A)} + \underbrace{\mathrm{Uni}(Z:B|A) + \mathrm{Syn}(Z:(A,B))}_{\mathrm{I}(Z;B|A)} \\ &= \underbrace{\mathrm{Uni}(Z:B|A) + \mathrm{Red}(Z:(A,B))}_{\mathrm{I}(Z;B)} + \underbrace{\mathrm{Uni}(Z:A|B) + \mathrm{Syn}(Z:(A,B))}_{\mathrm{I}(Z;A|B)}.\end{aligned}$$

Given three independent equations (1), (2) and (3) in four unknowns (the four PID terms), defining any one of the terms (e.g., $\mathrm{Uni}(Z:A|B)$) is sufficient to obtain the other three. For completeness, we include the definition of unique information from [59] (that also allows for estimation via convex optimization [62]) with the specific properties used in the proofs in Appendix B. To follow the paper, only an intuitive understanding is sufficient.

Definition 1 (Unique Information [59])**.** *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.,*

$$\Delta_p = \{Q \in \Delta : q(z,a){=}\Pr(Z{=}z, A{=}a) \text{ and } q(z,b){=}\Pr(Z{=}z, B{=}b)\}.$$

Then,

$$\mathrm{Uni}(Z:A|B) = \min_{Q \in \Delta_p} \mathrm{I}_Q(Z;A \mid B),$$

where $\mathrm{I}_Q(Z;A \mid B)$ *is the conditional mutual information when* (Z,A,B) *have joint distribution* Q*.*

The key intuition behind this definition is that the unique information should only depend on the marginal distribution of the pairs (Z,A) and (Z,B). This is motivated from an **operational** perspective that if A has unique information about Z (with respect to B), then there must be a situation where one can predict Z better using A than B (more details in [59, Section

2]). Therefore, all the joint distributions in the set Δ_p with the same marginals essentially have the same unique information, and the distribution Q^* that minimizes $\mathrm{I}_Q(Z; A \mid B)$ is the joint distribution that has no synergistic information leading to $\mathrm{I}_{Q^*}(Z; A \mid B) = \mathrm{Uni}(Z : A|B)$. Definition 1 also defines $\mathrm{Red}(Z : (A, B))$ and $\mathrm{Syn}(Z : (A, B))$ using (2) and (3).

B. System Model and Assumptions

Here, we will introduce our system model and assumptions. We start with an introduction to Structural Causal Model (SCM).

Definition 2 (Structural Causal Model: $\mathrm{SCM}(U, V, \mathcal{F})$ [50])**.** *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.*

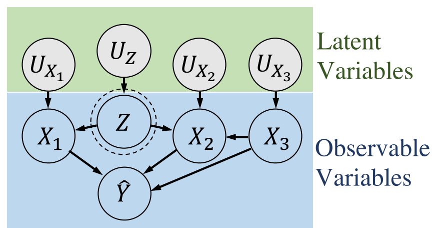


Fig. 2: An SCM with protected attribute Z, features $X = (X_1, X_2, X_3)$, and output $\hat{Y}$. Here X and $\hat{Y}$ are the observables, and U_Z and $U_X = (U_{X_1}, U_{X_2}, U_{X_3})$ are the latent social factors. Z does not have any parents in the SCM and $\hat{Y}$ is completely determined by $X = (X_1, X_2, X_3)$.

Our System Model: For our problem, consistent with several other works on fairness [18], [19], [21], the latent variables U represent possibly unknown social factors. The observables V consist of the protected attributes Z, the features X and the output $\hat{Y}$ (see Fig. 2). 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 (U_{X_j}) and its parents, which are again functions of their own latent variables and parents. Therefore, each X_j can also be written as $f_j(Z, U_X)$ for some deterministic $f_j(\cdot)$, where $U_X = U \backslash U_Z$ denotes the latent factors in U that do not cause Z (see a formal proof in [50, Proposition 6.3]). Here, $f_j(\cdot)$ may be constant in some of its arguments. This claim holds because the underlying graph is acyclic, and hence the structural assignments of the ancestors of X_j can be substituted recursively into one another until all observables except Z are substituted by latent variables. Also note that, $Z \perp\!\!\!\perp U_X$. A model takes X (which consists of critical features X_c and general features X_g) as its input and produces an output $\hat{Y}$ which is a deterministic function of X, i.e., $\hat{Y} = r(X)$ where X is itself a deterministic function of (Z, U_X). Therefore, $\hat{Y} = h(Z, U_X)$ for some deterministic function $h(\cdot)$.

Next, we introduce the concept of Counterfactual Causal Influence (CCI) ([18], [20], [63]–[67]), which will help us understand the well-known causal definition of fairness called *counterfactual fairness* [18].

Definition 3 (Counterfactual Causal Influence: $\mathrm{CCI}(Z \to \hat{Y})$)**.** *Consider the aforementioned system model. Let $\hat{Y} = h(Z, U_X)$ for some deterministic function $h(\cdot)$ where U_X are latent variables that do not cause Z in the true SCM. Then,*

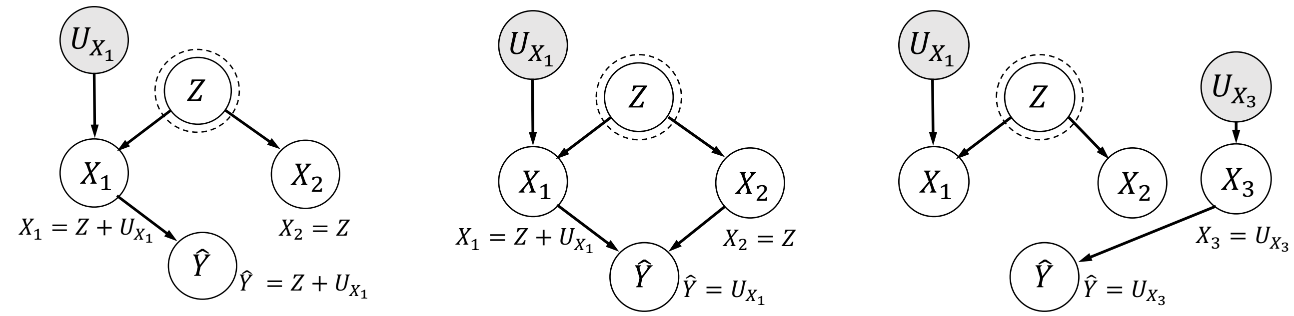
$$\mathrm{CCI}(Z \to \hat{Y}) = \mathbb{E}_{Z, Z', U_X}\left[|h(Z, U_X) - h(Z', U_X)|\right] \text{ where } Z', Z \text{ are i.i.d.}$$

Counterfactual causal influence quantifies the change in $\hat{Y} = h(Z, U_X)$ if we only vary Z while keeping the other latent factors (U_X) unchanged. A model is said to satisfy *counterfactual fairness* [18], [20] if and only if the output $\hat{Y}$ has no counterfactual causal influence of Z (we formally derive that $\mathrm{CCI}(Z \to \hat{Y}) = 0$ is equivalent to counterfactual fairness [18] in Lemma 6 in Appendix A-B). What this means is that a model is *counterfactually fair* if and only if the output $\hat{Y} = h(Z, U_X)$ does not change with Z while keeping the other latent factors (U_X) unchanged. It captures the intuitive notion that no virtual constituent or proxy of Z influences the output (inspired from the work on proxy-use [22]). In other words, $\hat{Y} \perp\!\!\!\perp Z | U_X$ (proved in Lemma 1), i.e.,

$$\Pr(\hat{Y} = y | Z = z, U_X = u_x) = \Pr(\hat{Y} = y | Z = z', U_X = u_x) \; \forall z, z', y, u_x.$$

This notion of fairness also leads us to propose an information-theoretic quantification of total bias (exempt and non-exempt) that is 0 if and only if the counterfactual causal influence of Z on $\hat{Y}$ is 0 (equivalence is demonstrated in Lemma 1 with the proof in Appendix A-A).

Definition 4 (Total Bias)**.** *The total bias in a model is defined as* $\mathrm{I}(Z; (\hat{Y}, U_X))$.



(a) Model is not counterfactually fair as $\hat{Y}$ has counterfactual causal influence of Z.
(b) Model is counterfactually fair after cancelling out the influence of Z from X_1.
(c) Model is counterfactually fair even though it uses an entirely unrelated feature.

Fig. 3: Examples for Counterfactual Fairness: Different models are used to make decisions on insurance premium corresponding to the same SCM with Z denoting the protected attribute, U_{X_1} denoting inherent tendency towards aggressive driving, $X_1 = Z + U_{X_1}$ denoting preference towards red cars, and X_3 denoting a feature not related to driving.

Notice that, $\mathrm{I}(Z;(\hat{Y},U_X)) = \mathrm{I}(Z;\hat{Y}|U_X) + \underbrace{\mathrm{I}(Z;U_X)}_{=0 \text{ since } Z \perp\!\!\!\perp U_X} = \mathrm{I}(Z;\hat{Y}|U_X)$.

Lemma 1 (Equivalences of CCI)**.** *Consider the aforementioned system model. Let $\hat{Y} = h(Z,U_X)$ for some deterministic function $h(\cdot)$ and $Z \perp\!\!\!\perp U_X$. Then, $\mathrm{CCI}(Z \to \hat{Y}) = 0$ if and only if $\mathrm{I}(Z;(\hat{Y},U_X)) = 0$.*

Remark 2 (Advantage of our Information-Theoretic Quantification)**.** *One might wonder why such an information-theoretic quantification of counterfactual causal influence (or, total bias) is necessary. The information-theoretic quantification of total bias enables analytical decomposition into exempt and non-exempt components that better satisfy our intuitive understanding. Our non-exempt bias intuitively attempts to capture whether discriminatory proxies are formed inside the black box model that cannot be entirely attributed to the critical features X_c. The decomposition of counterfactual causal influence (Definition 3) into exempt and non-exempt components is not straightforward. For instance, following the ideas of path-specific counterfactual fairness [21], 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, bias can also arise from synergistic information 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 be (see Counterexample 3). Purely causal measures can attribute such bias entirely to X_c. We contend that such synergistic information, if influencing the decision, must be included in the non-exempt component of bias because both X_c and X_g are contributors to the proxy. Information-theoretic equivalences of other existing notions of fairness, e.g., statistical parity, equalized odds, etc. have also been used in the broader literature on fairness [5], [7], [10], [35], [68].*

For a better understanding of counterfactual fairness, we now consider an example (inspired from [18]).

Example 2 (Counterfactual Fairness)**.** *Suppose a car insurance company makes its decisions about the insurance premium based on a feature X_1 which denotes your preference towards red [6] cars. In the SCM, this feature $X_1 = Z + U_{X_1}$ where Z denotes the protected attribute and U_{X_1} denotes the inherent tendency towards aggressive driving which is independent of Z. An output $\hat{Y} = X_1$ is not counterfactually fair because it has counterfactual causal influence of the protected attribute Z which is undesirable (Fig. 3a). The total bias $\mathrm{I}(Z;(\hat{Y},U_X))$ is also non-zero, capturing the intuitive notion that a proxy of Z influences the output. On the other hand, suppose the model now uses another feature $X_2 = Z$ and produces the output $\hat{Y} = X_1 - X_2 = U_{X_1}$. This model is now deemed counterfactually fair (Fig. 3b), and its total bias $\mathrm{I}(Z;(\hat{Y},U_X))$ is zero. No proxy of Z influences the output any longer.*

Remark 3 (Accuracy vs Counterfactual Fairness)**.** *The goals of fairness and accuracy on a given dataset are not always aligned [6], [15], [71], [72]. For instance, suppose the model in Example 2 takes decisions only based on a new feature $X_3 = U_{X_3}$ that is derived entirely from some latent factor that has got nothing to do with driving (see Fig. 3c). Such a model may be highly inaccurate but it is still counterfactually fair because it has no counterfactual causal influence of Z and does not cause any disparate impact with respect to Z. In this work, we will assume that a model has absolutely no bias (exempt or non-exempt) if and only if there is no counterfactual causal influence of Z on $\hat{Y}$. We will also run into some examples that might have lower accuracy, but it will be desirable that they are deemed fair if there is no counterfactual causal influence of Z.*

Next, we propose two definitions, namely, statistically visible bias and masked bias. Statistically visible bias is an information-theoretic quantification inspired from a well-known observational definition of fairness called *statistical parity* [2], [3].

[6]This example is inspired from [18]; also see [69], [70] to learn more about the debate on whether red cars are more prone to accidents, tickets, etc.

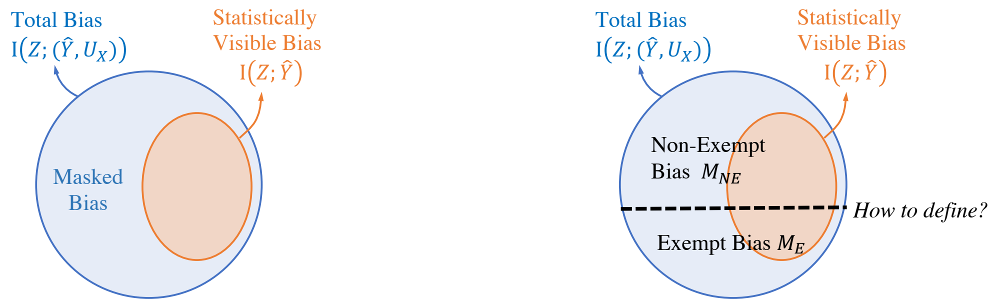


Fig. 4: Decomposition of Total Bias: (Left) Total bias (information-theoretic quantification of counterfactual causal influence) is shown in blue. The statistically visible bias and masked bias are two sub-components of the total bias. (Right) Our goal is to decompose the total bias into exempt and non-exempt components.

Definition 5 (Statistically Visible Bias)**.** *The statistically visible bias in a model is defined as* $\mathrm{I}(Z;\hat{Y})$.

Statistical parity deems a model fair if and only if $Z \perp\!\!\!\perp \hat{Y}$, i.e.,

$$\Pr(\hat{Y}=y|Z=z)=\Pr(\hat{Y}=y|Z=z') \quad \forall y,z,z'.$$

Thus, a model is said to be fair by statistical parity if and only if its statistically visible bias $\mathrm{I}(Z;\hat{Y})=0$.

Remark 4 (Statistical Parity vs Counterfactual Fairness)**.** *Statistical parity (or independence) does not imply absence of causal effects. E.g., consider* $\hat{Y}=Z\oplus U_X$ *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 *keeping all other sources of randomness in* $\hat{Y}$ *constant (i.e., fixing* $U_X=u_x$*), then* $\hat{Y}$ *also varies. This is, in fact, an example of masked bias, where* $\mathrm{I}(Z;\hat{Y})=0$*, but* Z *has counterfactual causal influence on* $\hat{Y}$*.*

Definition 6 (Masked Bias)**.** *The masked bias in a model is defined as* $\mathrm{I}(Z;(\hat{Y},U_X))-\mathrm{I}(Z;\hat{Y})$.

The masked bias is the difference between the total bias and the statistically visible bias. Notice that, $\mathrm{I}(Z;\hat{Y},U_X)-\mathrm{I}(Z;\hat{Y})=\mathrm{I}(Z;U_X\mid\hat{Y})$, implying that masked bias is non-negative. We will revisit masked bias in Section IV.

Goal: In this work, $\mathrm{I}(Z;(\hat{Y},U_X))$ will serve as our *information-theoretic quantification of the total bias (exempt and non-exempt)* as we discussed in Definition 4 (also recall Lemma 1 and Remark 2). Our *goal* is to appropriately decompose the total bias $\mathrm{I}(Z;(\hat{Y},U_X))$ into an exempt component (M_E) and a non-exempt component (M_{NE}), which can and cannot be explained by the critical features X_c (also see Fig. 4). Intuitively, the total bias captures the idea of a virtual constituent or proxy of Z that has a causal influence on the output $\hat{Y}$. We would like the exempt and non-exempt components of total bias to be able to capture and mathematically quantify our intuitive notion of what part of the virtual constituent or proxy can and cannot be attributed to the critical features X_c alone.

Before proceeding further, we also clarify our terminology here. We say that there is *no bias* when $\mathrm{I}(Z;\hat{Y},U_X)=0$. Alternately, we call *the bias to be exempt* if only the non-exempt component is 0, though $\mathrm{I}(Z;\hat{Y},U_X)$ may be zero or non-zero. Table II summarizes all the important notations to help follow the rest of the paper.

TABLE II: Summary of Notations

Symbol	Description	Observable or Not
X_c	Tuple of Critical features	Observable
X_g	Tuple of Non-critical or general features	Observable
X	Tuple of all input features (critical and general)	Observable
Z	Protected attribute (s)	Observable
U_X (Note that, $Z \perp\!\!\!\perp U_X$)	Tuple of latent social factors that do not cause Z	Not observable in general
$\hat{Y}=r(X)=h(Z,U_X)$	Model output	Observable

III. Main Result: Desirable Properties Leading to Our Proposed Measure of Non-Exempt Bias

In Section III-A, we first formally state the desirable properties that a measure of non-exempt bias (M_{NE}) should satisfy. These properties were only intuitively stated in Section I. Next, we introduce our proposed measure that satisfies all these properties (Theorem 1 in Section III-A). In Section III-B, we discuss in detail on how we arrive at these desirable properties through several examples, counterexamples and thought experiments, that helps us quantify our intuitive notion of non-exempt bias. In Section III-C, we examine measures in existing literature that have some provision for exemptions, namely, conditional statistical parity [57], path-specific counterfactual fairness [21] and justifiable fairness [56].

A. Main Result: Desirable Properties and a Proposed Measure Satisfying All of Them

We first state our set of desirable properties for any measure of non-exempt bias (M_{NE}). Firstly, if all the features are in the set X_g and there is no critical feature (i.e., $X_c = \phi$), then we would like M_{NE} to be equal to the total bias $\mathrm{I}(Z;(\hat{Y},U_X))$, i.e., the entire bias is non-exempt.

Property 1 (Absence of Exemptions). *If the set of critical features, $X_c = \phi$, then a measure M_{NE} should be equal to the total bias, i.e., $\mathrm{I}(Z;(\hat{Y},U_X))$.*

Next, for a fixed set of features and a fixed model $\hat{Y} = h(Z, U_X)$, it may also be desirable that the measure M_{NE} either decrease or stay the same if more features are removed from the set X_g and added to X_c.

Property 2 (Non-Increasing with More Exemptions). *For a fixed set of features X and a fixed model $\hat{Y} = h(Z, U_X)$, a measure M_{NE} should be non-increasing if a feature is removed from X_g and added to X_c.*

We require the measure M_{NE} to be 0 if all the features are in the exempt set X_c.

Property 3 (Complete Exemption). *M_{NE} should be 0 if all features are exempt, i.e., $X_c = X$ and $X_g = \phi$.*

Next, the measure should also avoid false positive conclusions, e.g., it should be 0 if the final model output has no counterfactual causal influence of Z, leading to the following property.

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

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 (more examples are provided in Section III-B to further motivate this property; see Counterexample 1, Counterexample 3, etc.).

Property 5 (Non-Exempt Statistically Visible Bias). *M_{NE} should be strictly greater than 0 if $\hat{Y}$ has any unique information about Z. Thus, $\mathrm{Uni}(Z:\hat{Y}|X_c) > 0$ should imply that $M_{NE} > 0$.*

However, statistical masking can sometimes prevent the entire non-exempt bias from exhibiting itself in $\mathrm{Uni}(Z:\hat{Y}|X_c)$ as demonstrated in the following example (recall Remark 4 in Section II). This example is inspired from [4], [18], [22].

Example 3 (Non-Exempt Masked Bias). *An ad for expensive housing is presented to white people ($Z = 1$) with income above a threshold ($U_{X_1} = 1$), and also to black people ($Z = 0$) with income below a threshold ($U_{X_1} = 0$) (while being largely irrelevant to the latter) with Z and U_{X_1} being i.i.d. Bern(½). Here, let $X_1 = U_{X_1}$ and $X_2 = Z$, where X_1 may or may not be critical but X_2 is definitely not critical, i.e., X_2 is a feature included in X_g. The model output is given by $\hat{Y} = Z \oplus U_{X_1}$ where $\oplus$ means XOR.*

It is evident that this model is unfair to high-income black people. This is a scenario where a virtual constituent Z is formed from X_g that is causally influencing the output (in a counterfactual sense; recall Definition 3 in Section II) but due to statistical masking, $\mathrm{I}(Z;\hat{Y}) = 0$. Since $\mathrm{Uni}(Z:\hat{Y}|X_c) \leq \mathrm{I}(Z;\hat{Y})$ (recall (2) in Section II-A and non-negativity of all PID terms), we have $\mathrm{Uni}(Z:\hat{Y}|X_c) = 0$ for this canonical example, showing that it fails to capture such "non-exempt masked bias."

Based on this example, one may feel that instead of examining the unique information about Z in $\hat{Y}$ alone, we should examine the unique information in $(\hat{Y}, U_X)$. The intuition is that the masked bias (or information) about Z in $\hat{Y}$ becomes exposed in $(\hat{Y}, U_X)$. Indeed, our total bias is also given by $\mathrm{I}(Z;\hat{Y},U_X)$ which captures the total (statistically visible and masked) bias, not captured by the statistically visible bias $\mathrm{I}(Z;\hat{Y})$ alone. This leads to a candidate measure $\mathrm{Uni}(Z:(\hat{Y},U_X)|X_c)$ to capture the entire non-exempt bias (statistically visible and masked). However, sometimes Z may already be statistically masked in X_c, e.g., $X_c = Z + U_{X_1}$ with $Z, U_{X_1} \sim i.i.d.$ Bern(½), and remain so in the final output, e.g., $\hat{Y} = X_c$ (Counterexample 4 in Section III-B). In such scenarios, it is desirable that the total bias be exempted and M_{NE} be 0 because the output is entirely derived from the critical feature. However, $\mathrm{Uni}(Z:(\hat{Y},U_X)|X_c)$ is greater than 0 here (Supporting Derivation 3 in Appendix C-B). This leads to the intuition that U_X can be divided into two subsets U_a and U_b (further functional generalizations discussed in Section VIII), where U_a consists of the latent factors that do not need to be accounted for as candidate masks, e.g., latent factors that do not influence $\hat{Y}$ at all, or those that already mask Z in X_c and continue to do so in the final output, as in the example above. On the other hand, U_b consists of the remaining latent factors, arising from both X_c and X_g, that need to be accounted for as candidate masks, e.g., those that possibly contribute to statistical masking "inside the black box model." This intuition leads to another desirable property (more examples are provided in Section III-B to further motivate this property; see Example 4, Counterexample 4, etc.).

Property 6 (Non-Exempt Masked Bias). *M_{NE} should be non-zero in Example 3, the canonical example of non-exempt masked bias. However, M_{NE} should be 0 if $(Z, U_a) - X_c - (\hat{Y}, U_b)$ form a Markov chain for some subsets $U_a, U_b \subseteq U_X$ such that $U_a = U_X \backslash U_b$.*

Intuitively, Properties 5 and 6 attempt to provide lower and upper bounds on a measure of non-exempt bias, i.e., it is desirable that $\mathrm{Uni}(Z:\hat{Y}|X_c) \leq M_{NE} \leq \min_{U_a,U_b \text{ s.t. } U_a=U_X\backslash U_b} \mathrm{I}((Z,U_a);(\hat{Y},U_b) \mid X_c)$. These two properties, in conjunction with the remaining four, lead to a novel measure of non-exempt bias that satisfies all of them (proved in Theorem 1).

Definition 7 (Non-Exempt Bias)**.** *Our proposed measure of non-exempt bias is given by:*

$$M^*_{NE} = \min_{U_a,U_b} \mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c) \textit{ such that } U_a = U_X\backslash U_b.$$

Theorem 1 (Properties)**.** *Properties 1-6 are satisfied by our proposed measure*

$$M^*_{NE} = \min_{U_a,U_b} \mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c) \textit{ such that } U_a = U_X\backslash U_b.$$

Proof Sketch: A detailed proof is provided in Appendix C-A. Here, we provide a brief proof sketch. For Property 1, we show that when $X_c=\phi$, we have $M^*_{NE} = \min_{U_a,U_b \text{ s.t. } U_a=U_X\backslash U_b} \mathrm{I}(Z,U_a;\hat{Y},U_b) = \mathrm{I}(Z;(\hat{Y},U_X))$. Property 2 is derived using a monotonicity property of unique information [73, Lemma 32]. For Property 3,

$$M^*_{NE} \leq \mathrm{Uni}(Z,U_X:\hat{Y}|X) \overset{(a)}{\leq} \mathrm{I}(Z,U_X;\hat{Y}|X) \overset{(b)}{=} 0, \tag{4}$$

where (a) holds because unique information is a component of conditional mutual information (see (3) in Section II-A) and (b) holds as $\hat{Y}$ is a deterministic function of X. For Property 4, note that

$$M^*_{NE} \leq \mathrm{Uni}(Z:\hat{Y},U_X|X_c) \leq \mathrm{I}(Z;(\hat{Y},U_X)), \tag{5}$$

where the last step holds as unique information is also a component of mutual information (see (2) in Section II-A). For Property 5, we show that $M^*_{NE} \geq \mathrm{Uni}(Z:\hat{Y}|X_c)$ using another monotonicity property of unique information [73, Lemma 31]. Lastly, for Property 6, we have $\mathrm{I}(Z,U_a;\hat{Y},U_b|X_c)=0$ for some U_a,U_b, implying that $\mathrm{Uni}(Z,U_a:\hat{Y},U_b|X_c)$ is also 0 for those U_a,U_b because unique information is a component of conditional mutual information (see (3) in Section II-A).

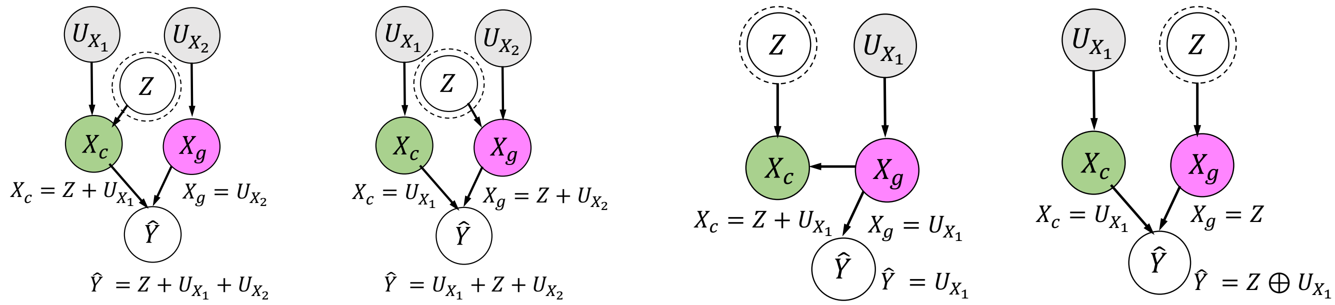
Remark 5 (On Exhaustive Set of Properties leading to a Unique Measure)**.** *We note that our properties do not quantify how exactly the non-exempt bias should "scale" when the measure is nonzero since they are only conditions on when this bias is nonzero, or on the monotonicity of this bias. Hence, these properties do not lead to a unique measure. Also, note that this is an issue with all measures of fairness in that they go to zero based on an intuitive notion of fairness but their exact scaling when they are non-zero is not unique. Neither do we claim that the proposed list of desirable properties (axioms) are exhaustive. In general, it is difficult to prove that a proposed set of properties (or, axioms) is exhaustive for a problem. E.g., Shannon established uniqueness on entropy with respect to* ***some*** *properties in [74] but the needs of the application can still drive the use of alternate measures. E.g. Renyi measures [25], [68], [75]–[77] have been found to be useful in security and privacy applications because they weigh outliers differently. Therefore, we believe, that there may be value in the measure not being unique so that it can be tuned to the needs of the application, as well as, motivate future work in this direction. Nonetheless, our properties do capture important aspects of the problem, e.g., non-exempt masked and non-exempt statistically visible biases, as discussed in Section IV.*

Next, we provide the rationale behind all our desirable properties using thought experiments.

B. Rationale Behind the Desirable Properties

It is desirable that our measure of *non-exempt bias* is able to capture and mathematically quantify our intuitive notion of a virtual constituent or proxy of Z being formed inside a given black box model that causally influences the output $\hat{Y}$ and that cannot be attributed to the critical features X_c alone. In order to propose such a quantification, we will examine several examples (thought experiments) that help us arrive at a set of desirable properties (axioms) that any measure of non-exempt bias should satisfy in order to be consistent with our intuitive notions. In these examples, we will explicitly note when a virtual constituent or proxy of Z is formed inside the model that influences the output, and whether it is entirely explainable by X_c.

Remark 6 (On Simplicity of Examples)**.** *We note that, at a first glance, our examples might seem quite simple, and real world models will only be more complex due to a mix of various causal and statistical relationships. These simple examples help us isolate many of these individual causal and statistical relationships, and examine them carefully. For instance, when both non-exempt masked and non-exempt statistically visible biases are present together, we are able to quantify both of them appropriately (discussed further in Section IV). Thus, developing an axiomatic understanding of such simple examples is an essential first step in understanding the complex interplay of various relationships in a real dataset. Indeed, examining toy examples (also called thought experiments) is a common practice in several works in existing fairness literature [18], [19], [36], [37], [56], some of which have also inspired our examples in this work. Furthermore, our quantification of non-exempt bias is*



(a) Hiring software engineers with coding skills (Example 4) (b) Discrimination in admissions (Example 5) (c) Counterfactually fair college admissions (Counterexample 1) (d) Non-exempt masked bias in housing ads (Counterexample 2)

Fig. 5: Thought experiments to understand total bias, exempt bias and non-exempt bias: In all the figures, Z denotes the protected attribute, e.g., gender, race, etc., and U_{X_1} denotes other latent social factors such that $Z \perp\!\!\!\perp U_{X_1}$.

not limited to black-box models alone, but also applies to "white-box" models [22], e.g., decision trees, linear classifiers, etc., and also to non-AI-based decisions as long as the decision can be represented as a deterministic function of the input features, i.e., $\hat{Y} = h(X)$.

Now, we move on to the rationale behind our desirable properties. The first three properties are more or less intuitive. For Property 1, if no feature is critical, it essentially means that all features are non-critical/general and no bias is exempt. So, we would like our measure of non-exempt bias to quantify the total bias given by $\mathrm{I}(Z; \hat{Y}, U_X)$. For Property 2, we argue that under a fixed model $\hat{Y} = h(Z, U_X)$, as the set of features designated as "critical" increases, a larger portion of the total bias is deemed exempt. Therefore the non-exempt bias is non-increasing (i.e., it decreases or stays the same). Ultimately, if all the features in a model are designated as critical, i.e., $X_c = X$, then the total bias becomes exempt, leading to M_{NE} being 0 (Property 3).

To arrive at the remaining desirable properties, we start out with examining two canonical examples that help us motivate the basic intuition behind *non-exempt bias*. These examples also help us understand the limitations of *statistical parity* [2], [3] and *equalized odds* [4] which are two popular measures of fairness that do not have provision for critical feature exemptions.

A Case against Statistical Parity:

As discussed in Section II, a model is deemed fair by statistical parity if $Z \perp\!\!\!\perp \hat{Y}$, i.e., $\mathrm{I}(Z; \hat{Y}) = 0$. However, the following example exposes some of its limitations.

Example 4 (Hiring Software Engineers with Coding Skills)**.** *Let* $X_c = Z + U_{X_1}$ *be the score in a coding test*[7] *and* $X_g = U_{X_2}$ *be a prior work experience score. Here the protected attribute* $Z \sim$ *Bern(½) denotes gender,* $U_{X_1} \sim$ *Bern(½) denotes inner ability to code and* $U_{X_2} \sim$ *Bern(½) denotes experience. An algorithm is deciding whether to hire software engineers based on a score* $\hat{Y} = Z + U_{X_1} + U_{X_2}$. *This is shown in Fig. 5a. Here + denotes addition (not to be confused with the binary OR).*

First notice that this model will be deemed *unfair* by both statistical parity and counterfactual fairness. Statistical parity is violated because Z and $\hat{Y}$ are not independent, i.e., the statistically visible bias $\mathrm{I}(Z; \hat{Y}) > 0$. Consequently, the total bias $\mathrm{I}(Z; (\hat{Y}, U_X))$ is also non-zero since $\mathrm{I}(Z; (\hat{Y}, U_X)) \geq \mathrm{I}(Z; \hat{Y}) > 0$, violating counterfactual fairness. However, for this example, the score in the coding test is a critical feature (bonafide requirement) for the job. Therefore, one may feel that any bias in $\hat{Y}$ that is explainable by the score in a coding test should be exempted. An attempt to ensure statistical parity for such an example, e.g., by reducing the importance (weight) of the critical feature in the decision making, violates the bonafide requirement of the job. Intuitively, even though the virtual constituent or proxy of Z, namely, $Z + U_{X_1}$, influences the output $\hat{Y}$, it is entirely explainable by X_c. Thus, for such an example, it is desirable that a measure of discrimination (non-exempt bias M_{NE}) be 0.

A Case against Equalized Odds:

Equalized odds [4], [10] is another popular measure of fairness that attempts to address this limitation of statistical parity by using the true labels (or scores) to represent the requirements of the job. Equalized odds states that a model is fair if

$$\Pr(\hat{Y} = y | Z = z, Y = \tilde{y}) = \Pr(\hat{Y} = y | Z = z', Y = \tilde{y}) \ \forall z, z', y, \tilde{y}.$$

This criterion is also equivalent to $\hat{Y} \perp\!\!\!\perp Z | Y$, or, $\mathrm{I}(Z; \hat{Y} \mid Y) = 0$. Indeed, in the previous example (Example 4), if the true scores already incorporate this critical requirement in them, e.g., $Y = Z + U_{X_1} + U_{X_2}$, then $\mathrm{I}(Z; \hat{Y} \mid Y) = 0$, and the model is deemed *fair* by equalized odds. While equalized odds is a reasonable quantification in scenarios where the true label (or score)

[7]The influence of Z on score in the SCM can arise due to various factors, e.g., historical lack of opportunities or sampling bias due to candidates of one protected group not applying enough etc. For instance, there may be a hidden node representing opportunity such that Z influences the score only though that hidden node, and the score becomes independent of Z given opportunity. We adopt a simplistic representation here for ease of understanding (also see [78]).

is indeed a justified representation of the job requirements, the measure $\mathrm{I}(Z;\hat{Y} \mid Y)$ has often criticized to be affected by label bias, as we demonstrate through this example.

Example 5 (Discrimination in Admissions)**.** *Let $X_c = U_{X_1}$ denote the score in a standardized test and $X_g = \begin{cases} U_X + 1, & Z = 0 \\ U_X, & Z = 1 \end{cases}$ denote the score from recommendation letters (biased). This can be rewritten as $X_g = Z(U_{X_2}+1) + (1-Z)U_{X_2} = Z + U_{X_2}$, where $Z \sim$ Bern(1/2) denotes gender, $U_{X_1} \sim$ Bern(1/2) denotes the latent ability and $U_{X_2} \sim$ Bern(1/2) denotes knowledge. Now suppose, the historic dataset has true selection scores given by $Y = U_{X_1} + Z + U_{X_2}$. This is shown in Fig. 5b.*

In this scenario, suppose we choose a perfect predictor, i.e., $\hat{Y} = Y = U_{X_1} + Z + U_{X_2}$. The perfect predictor always satisfies equalized odds because $\mathrm{I}(Z;\hat{Y} \mid Y) = 0$ if $\hat{Y} = Y$. However, if examined deeply, this model is propagating bias from recommendation letters, a non-critical feature, which is discriminatory and non-exempt. Intuitively, a virtual constituent or proxy of Z, i.e., $Z + U_{X_2}$, is being formed from X_g that is influencing the output $\hat{Y}$. For such an example[8], it is desirable that a measure of discrimination (non-exempt bias M_{NE}) is not zero.

Next, we start out with the aim of finding a suitable measure of non-exempt bias (M_{NE}) that takes the desirable values for both these canonical examples. Notice that, both these examples can be resolved by a notion of *conditional statistical parity* [57], which deems a model as fair if and only if $Z \perp\!\!\!\perp \hat{Y} | X_c$, i.e.,

$$\Pr(\hat{Y} = y | X_c = x_c, Z = z) = \Pr(\hat{Y} = y | X_c = x_c, Z = z') \;\; \forall y, x_c, z, z'.$$

This idea also connects with Simpson's paradox [50] which refers to a statistical trend that appears in several different groups of data but disappears or reverses when these groups are combined. In Example 4, Z and $\hat{Y}$ are not independent but they become so when conditioned on X_c, i.e., $\mathrm{I}(Z;\hat{Y}) > \mathrm{I}(Z;\hat{Y} \mid X_c)$. In Example 5, $\mathrm{I}(Z;\hat{Y}) < \mathrm{I}(Z;\hat{Y} \mid X_c)$. This notion of *conditional statistical parity* leads us to propose the following quantification of non-exempt bias (M_{NE}).

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

This measure resolves both Example 4 and 5. However, the following example exposes some of its limitations.

Counterexample 1 (Counterfactually Fair College Admissions)**.** *Let $Z \sim$ Bern(1/2) be the protected attribute, and let $U_X \sim \mathcal{N}(0, \sigma^2)$ be the latent ability of a student. Also, let $X_g = U_X$ be the score of a student in their high-school exam, but the score in an admission interview (critical feature) is given by $X_c = \begin{cases} U_X + 1, & Z = 0 \\ U_X, & Z = 1 \end{cases}$. This can be rewritten as $X_c = Z(U_X + 1) + (1-Z)U_X = Z + U_X$. Suppose the model for deciding admissions that maximizes accuracy is $\hat{Y} = X_g = U_X$. This is shown in Fig. 5c.*

Notice that, this model is deemed *fair* by counterfactual fairness because the total bias $\mathrm{I}(Z;(\hat{Y}, U_X)) = 0$. This means that the output $\hat{Y}$ has no counterfactual causal influence of Z. Even though the bias from X_c is legally exempt, the trained black-box model happens to base its decisions on another available non-critical feature that has no counterfactual causal influence of Z. Thus, there is no bias in the outcome $\hat{Y}$ (this is true even if the features in X_c were not exempt). Therefore, it is desirable that the non-exempt bias M_{NE} is also 0. This is also consistent with the intuition that here no virtual constituent or proxy of Z influences the output. However, the candidate measure $\mathrm{I}(Z;\hat{Y} \mid X_c) = \mathrm{I}(Z;U_X \mid Z + U_X)$ is non-zero here, leading to a false positive conclusion in detecting non-exempt bias.

Remark 7 (Cancellation of Paths)**.** *A similar situation arises if $X_c = Z + U_X$, $X_g = Z$ and $\hat{Y} = X_c - X_g = U_X$. Even though the bias from X_c may be exempt, the trained model ends up removing the counterfactual causal influence of Z from the decisions to make them counterfactually fair in a manner similar to the example of insurance premiums (recall Example 2 in Section II; also shown in Fig. 3b). The influences of Z along two different causal paths cancel each other in the final output, so that $\mathrm{CCI}(Z \rightarrow \hat{Y}) = 0$ (and, $\mathrm{I}(Z;(\hat{Y}, U_X)) = 0$). Since the total bias $\mathrm{I}(Z;(\hat{Y}, U_X)) = 0$, the question of non-exempt or exempt bias does not arise. However, the candidate measure $\mathrm{I}(Z;\hat{Y} \mid X_c)$ is non-zero here.*

This example also serves as a rationale for the property of zero influence, i.e., Property 4 which states that M_{NE} should be 0 if the total bias is 0. We aim to find a measure that resolves all of these examples (summarized in Fig. 5).

Motivation for Unique Information:

We notice that conditioning on the critical feature X_c can increase or decrease mutual information. In Example 4, we have $\mathrm{I}(Z;\hat{Y}) > 0$ but $\mathrm{I}(Z;\hat{Y} \mid X_c) = 0$. In Counterexample 1, $\mathrm{I}(Z;\hat{Y} \mid X_c) > 0$ but $\mathrm{I}(Z;\hat{Y}) = 0$. For both these cases, it is desirable that $M_{NE} = 0$. This motivates us to consider another candidate measure of non-exempt bias that is equal to the information-theoretic sub-volume of intersection between $\mathrm{I}(Z;\hat{Y})$ and $\mathrm{I}(Z;\hat{Y} \mid X_c)$ (recall Fig. 1b), that goes to 0 when any

[8]The example can be made more realistic if U_{X_1}, U_{X_2} are i.i.d. $\mathcal{N}(0,1)$. Now suppose, the historic dataset has true labels given by $Y = \mathrm{sgn}\left(Z + U_{X_1} + U_{X_2} - 0.5\right)$ which is binary. A perfect classifier $\hat{Y} = Y$, that satisfies equalized odds, is still discriminatory because it is influenced by Z in its decision, that is arising from recommendation letters, a non-critical feature.

one of them is 0. This is a quantity that is derived from the PID literature, and is called the *unique information* of Z in $\hat{Y}$ that is not present in X_c.

Candidate Measure 2. $M_{NE} = \mathrm{Uni}(Z : \hat{Y}|X_c)$.

We now show that this measure resolves the examples discussed so far, namely, Example 4 (Fig. 5a), Example 5 (Fig. 5b), Counterexample 1 (Fig. 5c) and a (similar) example in Remark 7. We start with Example 4 (the canonical example of hiring software engineers with coding skills), where $\hat{Y} = Z + U_{X_1} + U_{X_2}$ and $X_c = Z + U_{X_1}$. Recall that the total mutual information (statistically visible bias) can be decomposed as follows:

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z : \hat{Y}|X_c) + \mathrm{Red}(Z : (\hat{Y}, X_c)) \text{ (from (2) in Section II-A).}$$

For this example, we notice that even though $\mathrm{I}(Z;\hat{Y}) > 0$, we have $\mathrm{Uni}(Z : \hat{Y}|X_c) = 0$. This is because,

$$\mathrm{I}(Z;\hat{Y} \mid X_c) = \mathrm{Uni}(Z : \hat{Y}|X_c) + \mathrm{Syn}(Z : (\hat{Y}, X_c)) \text{ (from (3) in Section II-A),}$$

and $\mathrm{I}(Z;\hat{Y} \mid X_c) = 0$ for Example 4. In Example 4, the entire statistically visible bias $\mathrm{I}(Z;\hat{Y})$ is essentially redundant information between $\hat{Y}$ and X_c which is exempted.

Next, we revisit Example 5 ($\hat{Y} = U_{X_1} + Z + U_{X_2}$ and $X_c = U_{X_1}$) where it is intuitive that the measure of non-exempt bias should be non-zero. $\mathrm{Uni}(Z : \hat{Y}|X_c)$ is non-zero here (see Supporting Derivation 1 in Appendix C-B), consistent with our intuition. As a proof sketch, recall the tabular representation in Fig. 1b. $\mathrm{Red}(Z : (\hat{Y}, X_c))$ is the sub-volume of intersection between $\mathrm{I}(Z;X_c)$ and $\mathrm{I}(Z;\hat{Y})$, and hence goes to zero because $\mathrm{I}(Z;X_c) = 0$. This leads to $\mathrm{Uni}(Z : \hat{Y}|X_c) = \mathrm{I}(Z;\hat{Y})$ which is non-zero here.

Lastly, $\mathrm{Uni}(Z : \hat{Y}|X_c)$ is also 0 in Counterexample 1 (counterfactually fair college admissions) and the (similar) example of cancellation of paths in Remark 7. More importantly, we note that, while conditional mutual information $\mathrm{I}(Z;\hat{Y} \mid X_c)$ may be non-zero even if the the total bias or counterfactual causal influence is 0 (as in Counterexample 1), unique information is not. *In Lemma 13 in Appendix B, we show that* $\mathrm{Uni}(Z : \hat{Y}|X_c)$ *is always zero if the total bias or counterfactual causal influence is* 0*, i.e.,* $\mathrm{I}(Z;(\hat{Y}, U_X)) = 0$. In fact, $\mathrm{Uni}(Z : \hat{Y}|X_c)$ is a sub-volume or component of the previous candidate measure $\mathrm{I}(Z;\hat{Y} \mid X_c)$, that is guaranteed to be 0 if the total bias is zero.

These examples serve as our rationale for the property of non-exempt statistically visible bias, i.e., Property 5 which states that M_{NE} should be 0 if $\mathrm{Uni}(Z : \hat{Y}|X_c) > 0$. In fact, $\mathrm{Uni}(Z : \hat{Y}|X_c)$ is the information-theoretic sub-volume (intersection) between $\mathrm{I}(Z;\hat{Y})$ and $\mathrm{I}(Z;\hat{Y} \mid X_c)$, that captures the *non-exempt* part of the statistically visible bias $\mathrm{I}(Z;\hat{Y})$.

$\mathrm{Uni}(Z : \hat{Y}|X_c)$, however, is not sufficient as a candidate measure as it fails to capture *non-exempt masked bias*, as we will demonstrate in Counterexample 2. Thus, Property 5 is a one-way implication only, i.e., sometimes M_{NE} may still need to be non-zero even when $\mathrm{Uni}(Z : \hat{Y}|X_c) = 0$. Property 5 only captures the *non-exempt statistically visible bias* that cannot be accounted for by X_c alone.

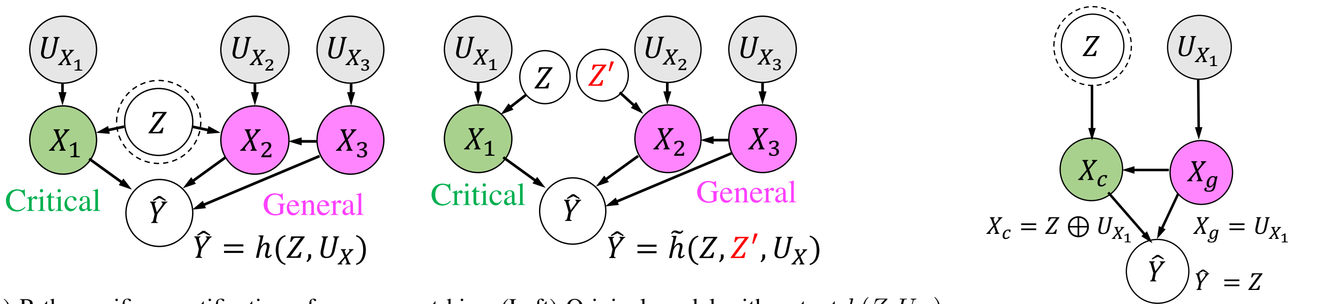
Counterexample 2 (Non-Exempt Masked Bias in Housing Ads; A Special Case of Example 3 in Section III-A)**.** *Let* $X_c = U_{X_1}$ *and* $X_g = Z$ *where the protected attribute* $Z \sim$ *Bern(½) denotes race and* $U_{X_1} \sim$ *Bern(½) denotes whether the income is above a threshold. The model decides to show an expensive housing ad based on* $\hat{Y} = Z \oplus U_{X_1}$ *(same as Example 3 in Section III-A with* $X_c = X_1$*). This is shown in Fig. 5d.*

This model will be deemed *fair* by statistical parity because $Z \perp\!\!\!\perp \hat{Y}$, *i.e.*, the statistically visible bias $\mathrm{I}(Z;\hat{Y}) = 0$. However, the model racially discriminates against half of the population (high-income people with $Z = 0$) for whom the housing ad is relevant. This is also supported by the fact that that the total bias $\mathrm{I}(Z;(\hat{Y}, U_X)) > 0$. Intuitively, here a virtual constituent or proxy (Z) is formed inside the black box model that influences the output and that is derived entirely from X_g. For such an example, it is desirable that the non-exempt bias M_{NE} should not be 0. In fact, this example demonstrates that there may be non-exempt bias even when the statistically visible bias $\mathrm{I}(Z;\hat{Y}) = 0$ but $\mathrm{I}(Z;\hat{Y} \mid X_c) > 0$. Here, $\mathrm{Uni}(Z : \hat{Y}|X_c)$ fails to capture the masked bias because it has to be zero whenever $\mathrm{I}(Z;\hat{Y}) = 0$ (using (2) in Section II-A).

One commonality that we notice in the examples so far (see Fig. 5) is that whenever it is desirable that M_{NE} be zero, either there is no counterfactual causal influence of Z on $\hat{Y}$ (i.e., $\mathrm{CCI}(Z \to \hat{Y}) = 0$) or the influence of Z on $\hat{Y}$ has propagated *only* along paths that pass through X_c. In scenarios where $\mathrm{CCI}(Z \to \hat{Y}) \neq 0$, we will define another candidate measure of non-exempt bias that is inspired from the notion of *path-specific counterfactual fairness* [21] (also see [18], [19]). The next candidate measure for quantifying non-exempt bias is a causal, path-specific quantification by varying Z only along the direct paths through X_g and comparing if it causes any change in the model output (also see Fig. 6a).

Candidate Measure 3. *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* $\hat{Y} = \tilde{h}(Z, Z', U_X)$ *in the new causal graph. A candidate measure is* $M_{NE} = \mathbb{E}_{Z,Z',U_X}\left[|h(Z, U_X) - \tilde{h}(Z, Z', U_X)|\right]$.

This measure, when used in conjunction with $\mathrm{CCI}(Z \to \hat{Y}) = 0$, resolves the examples so far (see Fig. 5). For Example 4, it is zero and for Example 5, it is non-zero, as desired. For Counterexample 1, $\mathrm{CCI}(Z \to \hat{Y}) = 0$, and hence there is no need



(a) Path-specific quantification of non-exempt bias: (Left) Original model with output $h(Z, U_X)$. (Right) Z is varied to Z' along direct paths through X_g resulting in output $\tilde{h}(Z, Z', U_X)$. Candidate measure 3 quantifies the expected value of the change in output due to path-specific variation in Z.

(b) Discrimination by unmasking (Counterexample 3)

Fig. 6: Path-specific quantification of non-exempt bias (Candidate measure 3) and its counterexample

for a path-specific examination. Lastly, for the example of non-exempt masked bias (Counterexample 2), *this measure is* 0 *in spite of the statistically visible bias* $\mathrm{I}(Z; \hat{Y})$ *being* 0. However, the following example exposes some of its limitations.

Counterexample 3 (Discrimination by Unmasking)**.** *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 $\hat{Y} = X_c \oplus X_g = Z$. This is shown in Fig. 6b.*

The bias in this example will be deemed *exempt* by a causal path-specific examination. However, this model has statistically visible bias ($\mathrm{I}(Z; \hat{Y}) > 0$) that cannot be attributed to X_c alone. Following the PID literature, here X_c and X_g have synergistic information about Z that ultimately appears in $\hat{Y}$ which in itself is the virtual constituent or proxy of Z being formed in this model. This synergistic information cannot be attributed to X_c alone because $\mathrm{I}(Z; X_c) = 0$. This is further supported by the argument that X_g 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 bias. Thus, M_{NE} should be greater than 0. This is further supported by the fact that here $\mathrm{Uni}(Z : \hat{Y} | X_c) > 0$ (Supporting Derivation 2 in Appendix C-B) because it is this "joint" information about Z in (X_c, X_g) that ultimately appears in $\hat{Y}$ that cannot be attributed to X_c alone.

Next, our question of interest is: *how to arrive at a property and a measure that correctly captures the intuition of non-exempt masked bias?* Let us revisit the candidate measure $\mathrm{I}(Z; \hat{Y} \mid X_c)$. This measure resolves all the examples so far (see Fig. 5 and Fig. 6b) except giving a false positive conclusion in Counterexample 1. Notice that, $\mathrm{I}(Z; \hat{Y} \mid X_c)$ is zero if and only if $Z - X_c - \hat{Y}$ form a Markov chain. While the Markov chain $Z - X_c - \hat{Y}$ may not always hold even when it is desirable for M_{NE} to be zero as in Counterexample 1, we have seen that in all the examples so far where the Markov chain $Z - X_c - \hat{Y}$ holds, it has been desirable that M_{NE} be zero (possible one-way implication). Assuming that the Markov chain $Z - X_c - \hat{Y}$ is a sufficient condition for M_{NE} to be zero, we proposed the following property of non-exempt masked bias in our prior work [1].

M_{NE} should be non-zero in the example of non-exempt masked bias, i.e., Counterexample 2 (special case of Example 3 with $X_c = U_{X_1}$, $X_g = Z$, and $\hat{Y} = Z \oplus U_{X_1}$) even if $\mathrm{I}(Z; \hat{Y}) = 0$*. But, M_{NE} should be* 0 *if the Markov chain $Z - X_c - \hat{Y}$ holds.*

Remark 8 (Relation to our prior work [1])**.** *In our prior work [1], this property, in conjunction with Properties 3, 4, and 5, leads to a measure that quantifies only a sub-volume of* $\mathrm{I}(Z; \hat{Y} \mid X_c)$ *that no longer gives false positive conclusion in Counterexample 1 while still resolving all the other examples. The measure proposed in [1] is essentially the information-theoretic sub-volume of the intersection between* $\mathrm{I}(Z; \hat{Y} \mid X_c)$ *and total bias* $\mathrm{I}(Z; (\hat{Y}, U_X))$*, which goes to* 0 *whenever either of them is* 0 *(details are provided in Appendix C-C)*[9].

The property of non-exempt masked bias stated in [1] is built on the rationale that in the example of non-exempt masked bias in housing ads (Example 3 where $\hat{Y} = Z \oplus U_{X_1}$), instead of U_{X_1} being the income, if U_{X_1} is a random coin flip used to randomize the race, then this scenario may not necessarily be regarded as non-exempt. Then, we would have $X_c = \phi$ and $X_g = (Z, U_{X_1})$, and the Markov chain $Z - X_c - \hat{Y}$ would hold, deeming this example as *exempt*. In [1], the goal was to only account for non-exempt masked bias in M_{NE} when the "mask" is either a critical feature or arises exclusively from the critical features, e.g., Counterexample 2 (a special case of Example 3 with $X_c = U_{X_1}$, $X_g = Z$ and $\hat{Y} = Z \oplus U_{X_1}$) while any mask from the non-critical or general features were viewed more like these random coin flips. But what if the user wishes to also account for masked bias if the mask is arising from X_g as well, as demonstrated in the following example.

Example 6 (Non-Exempt Masked Bias in Housing Ads; A Special Case of Example 3 in Section III-A)**.** *Let $X_c = \phi$ and $X_g = (Z, U_{X_1})$ with $Z, U_{X_1} \sim$ Bern(½) and $\hat{Y} = Z \oplus U_{X_1}$ (U_{X_1} is still income like in Example 3 but is not a critical feature).*

[9]One might also wonder why a measure of the form of a product, i.e., $M_{NE} = \mathrm{I}(Z; \hat{Y} \mid X_c) \times \mathrm{I}(Z; (\hat{Y}, U_X))$ does not work instead. We discuss a counterexample for such a product measure in [1] that we also include in Appendix C-C here for completeness.

Example 6 will be deemed *exempt* by [1] because the Markov chain $Z - X_c - \hat{Y}$ holds. Consequently, the measure proposed in [1] will also be 0 because it is a non-negative sub-component of $\mathrm{I}(Z;\hat{Y}|X_c)$. However, here the virtual constituent or proxy Z is arising from X_g and is being masked by another feature of X_g, i.e., U_{X_1}. If U_{X_1} denotes income and $\hat{Y}$ denotes the decision of showing housing ads, then the model is again unfair to high-income people of one race. This argument is also supported by the fact that the total bias is non-zero (not counterfactually fair). Since $X_c = \phi$, no bias is exempt, and a measure of non-exempt bias should ideally capture the total bias in this model.

In this work, we focus on defining an alternate criterion (modification of the property of non-exempt masked bias in [1]) that can capture non-exempt masked bias irrespective of whether the "mask" arises from the critical or non-critical features. What this means is that any scenario deemed exempt by the property of non-exempt masked bias in [1] will also be deemed exempt by our modified property[10] but it is desirable that our modified property also accounts for additional scenarios, such as Example 6, that are deemed exempt by the former property even though intuitively, it may not be reasonable to do so.

Example 6 reveals that the Markov chain $Z - X_c - \hat{Y}$ holding may not always imply that there is no non-exempt bias. In the output $\hat{Y}$, the constituent Z may be masked by some elements of U_X, e.g., $U_b \subseteq U_X$ that could arise from both X_c and X_g. Based on the examples of masked bias, one might be tempted to examine the information about Z present in $(\hat{Y}, U_X)$ instead of $\hat{Y}$ alone, e.g., examine the following Markov chain: $Z - X_c - (\hat{Y}, U_X)$ instead of $Z - X_c - \hat{Y}$. Motivated by this idea, we now consider another candidate measure $\mathrm{Uni}(Z:(\hat{Y}, U_X)|X_c)$, i.e., the unique information about Z present jointly in $(\hat{Y}, U_X)$ (instead of $\hat{Y}$ alone), that is not present in X_c. Again note that, we consider unique information instead of conditional mutual information because of two reasons: (i) The conditional mutual information $\mathrm{I}(Z;(\hat{Y}, U_X) \mid X_c)$ corresponding to this Markov chain is greater than or equal to $\mathrm{I}(Z;\hat{Y} \mid X_c)$, and therefore will again lead to a false conclusion about detecting non-exempt bias in Counterexample 1 where the total bias is 0. On the other hand, we show in Lemma 13 in Appendix B that $\mathrm{Uni}(Z:(\hat{Y}, U_X)|X_c)$ is less than or equal to the total bias $\mathrm{I}(Z;(\hat{Y}, U_X))$, and thus can never be non-zero when the total bias is 0. (ii) Unique information, being a sub-volume of conditional mutual information, goes to zero whenever the corresponding conditional mutual information is zero (i.e., the Markov chain holds).

Candidate Measure 4. $M_{NE} = \mathrm{Uni}(Z:(\hat{Y}, U_X)|X_c)$.

While this measure resolves all the examples so far (Fig. 5 and Fig. 6b), the following example exposes some of its limitations.

Counterexample 4 (Complete Exemption)**.** *Suppose that $X_c = Z + U_{X_1}$, $X_g = U_{X_2}$, and $\hat{Y} = X_c = Z + U_{X_1}$ where $Z, U_{X_1}, U_{X_2} \sim$ i.i.d. Bern(½).*

Here, $\hat{Y}$ is entirely derived from X_c which is the critical feature. Thus, the total bias is exempt (also notice the Markov chain $Z - X_c - \hat{Y}$). However, Candidate Measure 4 can be shown to be non-zero here using the properties of unique information in Definition 1 (Supporting Derivation 3 in Appendix C-B). The key step is that Z can be obtained entirely from deterministic local operations on $(\hat{Y}, U_X)$ and therefore $\mathrm{Uni}(Z:(\hat{Y}, U_X)|X_c) \geq \mathrm{Uni}(Z:Z|X_c)$ using a monotonicity property of unique information. Because unique information is a sub-volume of conditional mutual information, it being non-zero also implies that the Markov chain $Z - X_c - (\hat{Y}, U_X)$ does not hold.

Ideally, we would like a property and a measure that captures the intuition in this example. The main intuition is that Z was already masked by U_{X_1} in X_c, and it remained so in the output $\hat{Y}$ (unlike Counterexample 3 where the mask on Z in X_c was tampered inside the model). *In this example, we do not need to account for the masking of Z by U_{X_1} because it did not happen inside the black-box model, and neither did it get tampered in any manner inside the model.* Notice that, the Markov chain $Z - X_c - (\hat{Y}, U_b)$ holds *if* $U_b \subseteq U_X \backslash U_{X_1}$, i.e., if U_{X_1} is removed from the candidate masks that one needs to account for. This motivates the possibility that one might be able to split the set of latent factors U_X (see Sec. VIII for a functional generalization) into two parts (say U_a and U_b) such that U_a consists of all the latent factors that either do not influence $\hat{Y}$ at all, or already mask the constituent Z in X_c and remain untampered in $\hat{Y}$. The other part U_b corresponds to candidate masks that one might want to account for, e.g., masking effects that happened inside the black-box model (recall Example 4 in Fig. 5a), or if Z was already masked by some latent factors inside X_g (recall Example 5 in Fig. 5b). To understand this better, let us revisit Example 4 (canonical example of hiring software engineers with coding skills) again.

Intuitively, the total bias in this example is exempt because Z was already masked by U_{X_1} in X_c, and the mask remained untampered in the final output with only additional independent masks added inside the black-box model. The criterion $Z - X_c - (\hat{Y}, U_X)$ does not hold here. But, interestingly, the following Markov chain does hold: $Z - X_c - (\hat{Y}, U_{X_2})$. Here, again one might choose to only account for the masking by U_{X_2} because it is happening inside the black-box model. Thus, here we can let $U_a = U_{X_1}$, and $U_b = U_{X_2}$. Interestingly, we notice that $(Z, U_a) - X_c - \hat{Y}$ also form a Markov chain because U_a consists of only the latent factors that already mask Z in X_c and remain untampered in the final output $\hat{Y}$, e.g., U_{X_1} in Example 4. One might therefore consider a candidate measure of non-exempt bias given by $\mathrm{Uni}(Z:(\hat{Y}, U_b)|X_c)$ where $U_b \subseteq U_X$ is an appropriately defined subset of the candidate masks that one wishes to account. This measure, as stated, is

[10] We show in Lemma 2 that the Markov chain in our modified property, i.e., $(Z, U_a) - X_c - (\hat{Y}, U_b)$ also implies $Z - X_c - \hat{Y}$, but the opposite implication is not true.

somewhat ill-defined because the set U_b also depends on the choice of the critical features, and hence would need to be identified for different choices of X_c for the same mathematical scenario.

To arrive at a measure that is more precisely defined, we now examine an alternate criterion, namely, a Markov chain of the form $(Z, U_a) - X_c - (\hat{Y}, U_b)$ that in fact implies both the criterion $(Z, U_a) - X_c - \hat{Y}$ and $Z - X_c - (\hat{Y}, U_b)$ (see Lemma 2 with proof in Appendix C-A). This is motivated from the previous example, where neither $Z - X_c - (\hat{Y}, U_X)$ nor $(Z, U_X) - X_c - \hat{Y}$ hold, but $(Z, U_{X_1}) - X_c - (\hat{Y}, U_{X_2})$ does. This leads us to propose the following criterion for any measure of non-exempt bias (M_{NE}) that also serves as our main rationale for Property 6.
M_{NE} should be 0 if $(Z, U_a) - X_c - (\hat{Y}, U_b)$ form a Markov chain for some subsets $U_a, U_b \subseteq U_X$ such that $U_a = U_X \backslash U_b$.

Lemma 2. *The Markov chain $(Z, U_a) - X_c - (\hat{Y}, U_b)$ implies that the following Markov chains also hold: (i) $Z - X_c - \hat{Y}$; (ii) $(Z, U_a) - X_c - \hat{Y}$; and (ii) $Z - X_c - (\hat{Y}, U_b)$.*

The Markov chain $(Z, U_a) - X_c - (\hat{Y}, U_b)$ holding implies M_{NE}=0, but the Markov chain not holding for all U_a, U_b such that $U_a = U_X \backslash U_b$ does not necessarily imply that $M_{NE} \neq 0$. This criterion $(Z, U_a) - X_c - (\hat{Y}, U_b)$ implying $M_{NE} = 0$ only attempts to provide an upper bound on M_{NE}, i.e., it is desirable that $M_{NE} \leq \min_{U_a, U_b \text{ s.t. } U_a = U_X \backslash U_b} \mathrm{I}((Z, U_a); (\hat{Y}, U_b) \mid X_c)$ such that $U_a = U_X \backslash U_b$. The measure $\min_{U_a, U_b \text{ s.t. } U_a = U_X \backslash U_b} \mathrm{I}((Z, U_a); (\hat{Y}, U_b) \mid X_c)$ does not suffice in itself as a measure of non-exempt bias because it again does not satisfy Property 4. To see this, notice that $\min_{U_a, U_b \text{ s.t. } U_a = U_X \backslash U_b} \mathrm{I}((Z, U_a); (\hat{Y}, U_b) \mid X_c) \geq \mathrm{I}(Z; \hat{Y} \mid X_c)$ (see proof of Lemma 2), and thus, it also gives a false positive conclusion about non-exempt bias in Counterexample 1 (counterfactually fair college admissions). Property 6, in conjunction with all the other properties, ultimately lead us to our proposed measure of non-exempt bias, given by:

$$M_{NE}^* = \min_{U_a, U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) \text{ such that } U_a = U_X \backslash U_b.$$

This proposed measure satisfies all of the six desirable properties (see proof of Theorem 1 in Appendix C-A). To develop intuition on what this measure captures, we will now discuss how this measure resolves all of the examples in this work. We group "similar" examples together.

Scenarios where total bias $\mathrm{I}(Z; (\hat{Y}, U_X))$ is zero: This applies to Counterexample 1 and the related example in Remark 7. Because $\min_{U_a, U_b \text{ s.t. } U_a = U_X \backslash U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) \leq \mathrm{Uni}(Z : (\hat{Y}, U_X) | X_c) \leq \mathrm{I}(Z; (\hat{Y}, U_X))$ (see proof of Theorem 1 in Appendix C-A), it satisfies Property 4 and goes to 0 whenever total bias is 0.

Scenarios where Z is already masked in X_c and remains so in the output (with or without additional independent masks): This applies to Example 4 and Counterexample 4. We will only discuss Example 4 here since it can explain the other one as well. We will examine the value of $\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c)$ for different choices of $U_a \subseteq U_X$ to find the minimum. First notice that, if $U_a = \phi$ (and $U_b = U_X$), we have

$$\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) = \mathrm{Uni}(Z : (\hat{Y}, U_X) | X_c) \overset{(a)}{\geq} \mathrm{Uni}(Z : Z | X_c) > 0$$

(see Supporting Derivation 3 in Appendix C-B; (a) holds from a monotonicity property of unique information because Z can be obtained from deterministic local operations on $(\hat{Y}, U_X)$). This is in agreement with the intuition that U_{X_1} should not belong to the set of candidate masks (U_b) that need to be accounted for. Next, if $U_a = U_{X_1}$ (and $U_b = U_{X_2}$), we have $\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) = 0$ (implied from the Markov chain $(Z, U_{X_1}) - X_c - (\hat{Y}, U_{X_2})$). Since unique information is non-negative, we therefore have $\min_{U_a, U_b \text{ s.t. } U_a = U_X \backslash U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) = 0$. In essence, the pair (U_a^*, U_b^*) that minimizes $\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c)$ is such that $U_a^* = U_{X_1}$, and the candidate masks that need to be accounted for, i.e., $U_b^* = U_{X_2}$.

Now, what happens to the value of $\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c)$ if the accountable mask U_{X_2} is instead in U_a? We have

$$\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) \overset{(a)}{\geq} \mathrm{Uni}(U_{X_2} : \hat{Y} | X_c) \overset{(b)}{=} \mathrm{I}(U_{X_2}; \hat{Y}),$$

which is strictly greater than 0. This agrees with the intuition that U_{X_2} should belong to the candidate set of masks that one should account for (U_b). Here (a) holds using two monotonicity properties of unique information (see Properties 10 and 9 in Appendix B) and (b) holds because $\mathrm{I}(U_{X_2}; X_c) = 0$, leading to $\mathrm{Red}(U_{X_2} : (\hat{Y}, X_c)) = 0$.

Scenarios where non-exempt statistically visible bias is present, i.e., $\mathrm{Uni}(Z : \hat{Y} | X_c) > 0$: This applies to Example 5 and Counterexample 3. Because $\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) \geq \mathrm{Uni}(Z : \hat{Y} | X_c)$ (see proof of Theorem 1 in Appendix C-A), our proposed M_{NE}^* satisfies Property 5, and is thus non-zero whenever $\mathrm{Uni}(Z : \hat{Y} | X_c) > 0$.

Scenarios where non-exempt masked bias is present: This applies to Counterexample 2 and Example 6 (special cases of the canonical example of non-exempt masked bias, i.e., Example 3). In the proof of Theorem 1 in Appendix C-A, we show that the proposed measure satisfies Property 6 (non-exempt masked bias), and is thus non-zero for these canonical examples of non-exempt masked bias.

We note that Example 5 is an interesting case where both non-exempt statistically visible bias and non-exempt masked bias are present. Here, M^*_{NE} is strictly greater than the non-exempt statistically visible bias $(\mathrm{Uni}(Z : \hat{Y}|X_c))$, and this difference can be interpreted as a quantification of the non-exempt masked bias. First notice that,

$$\mathrm{Uni}(Z : \hat{Y}|X_c) \overset{(a)}{=} \mathrm{I}(Z; \hat{Y}) = \mathrm{H}(Z) - \mathrm{H}(Z|\hat{Y}) = \mathrm{H}(Z) - \mathrm{H}(Z|U_{X_1} + Z + U_{X_2}) = 1 - \frac{3}{4} h_b(1/3) \text{ bits.} \tag{6}$$

The full derivation is in Supporting Derivation 4 in Appendix C-B. Here $h_b(\cdot)$ is the binary entropy function [79] given by $h_b(p) = -p\log_2(p) - (1-p)\log_2(1-p)$ and (a) holds because $\mathrm{I}(Z; U_{X_1}) = 0$, implying $\mathrm{Red}(Z : (\hat{Y}, U_{X_1})) = 0$ as well. Now, we will examine the value of $\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b)|X_c)$ for different choices of U_a to find the minimum. The full derivation for all of these cases is in Supporting Derivation 4 in Appendix C-B. Here, we only mention the key step. Let $U_a = \phi$ (and $U_b = U_X$). Then,

$$\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b)|X_c) = \mathrm{Uni}(Z : (\hat{Y}, U_{X_1}, U_{X_2})|U_{X_1}) \overset{(a)}{=} \mathrm{I}(Z; U_{X_1} + Z + U_{X_2}, U_{X_1}, U_{X_2}) = 1 \text{ bit.} \tag{7}$$

Here (a) holds again because $\mathrm{I}(Z; U_{X_1}) = 0$, implying the redundant information is 0 as well (using (2) in Section II-A). Next, for $U_a = U_{X_2}$ (and $U_b = U_{X_1}$), we have,

$$\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b)|X_c) = \mathrm{Uni}((Z, U_{X_2}) : (\hat{Y}, U_{X_1})|U_{X_1}) \overset{(a)}{=} \mathrm{I}((Z, U_{X_2}); (\hat{Y}, U_{X_1})) = 3/2 \text{ bit.} \tag{8}$$

Here (a) holds again because $\mathrm{I}((Z, U_{X_2}); U_{X_1}) = 0$, implying the redundant information is 0 as well. Next, for $U_a = U_{X_1}$ (and $U_b = U_{X_2}$), we have,

$$\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b)|X_c) = \mathrm{Uni}((Z, U_{X_1}) : (\hat{Y}, U_{X_2})|U_{X_1}) \overset{(b)}{=} \mathrm{I}((Z, U_{X_1}); (\hat{Y}, U_{X_2}) \mid U_{X_1}) = 1 \text{ bit.} \tag{9}$$

Here (b) holds because $\mathrm{Syn}((Z, U_{X_1}) : (A, B)) = 0$ if one of the terms A or B is a deterministic function of (Z, U_{X_1}) (using Lemma 14 in Appendix B) and hence unique information becomes equal to the conditional mutual information (see (3) in Section II-A). Lastly, for $U_a = U_X$ (and $U_b = \phi$), we have,

$$\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b)|X_c) = \mathrm{Uni}((Z, U_{X_1}, U_{X_2}) : \hat{Y}|U_{X_1}) \overset{(b)}{=} \mathrm{I}((Z, U_{X_1}, U_{X_2}); \hat{Y} \mid U_{X_1}) = 3/2 \text{ bit.} \tag{10}$$

Here (b) holds again using Lemma 14 in Appendix B. Thus, we obtain that,

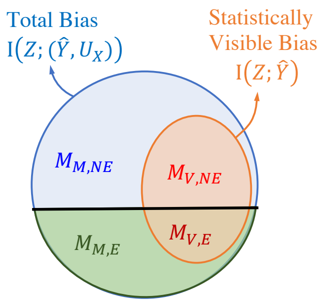
$$M^*_{NE} = \min_{U_a, U_b \text{ s.t. } U_a = U_X \setminus U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b)|X_c) = 1 \text{ bit,}$$

which is strictly greater than $\mathrm{Uni}(Z : \hat{Y}|X_c) = 1 - \frac{3}{4} h_b(1/3)$ bits, accounting for both non-exempt statistically visible and non-exempt masked biases.

As noted in Remark 5, our properties are insufficient to arrive at a unique functional form for the measure of non-exempt bias. It is easiest to understand this issue by contrasting it with Shannon's discussion on entropy as a measure for uncertainty. First, we do not have a counterpart of "additivity" of entropy (see Property 3 in Section 6 of [74]) which allows Shannon to arrive at the logarithmic scaling in entropy. Second, we also do not provide an operational meaning for this measure (such as that provided by the lossless source coding theorem for entropy [79]), which further supports the logarithmic scaling. This is a direction of meaningful future work (further functional generalizations discussed in Section VIII). We note that this is the case with almost all existing measures of fairness (with the notable exceptions of [25], [68], [76]). Exploring more deeply the desirable attributes of the influence of a virtual constituent or proxy of Z that influences the model output and that cannot be attributed to the critical features X_c alone (inspired from the work on proxy-use [22]) could be a starting point towards deriving an exact operational meaning for our proposed measure. Nonetheless, our measure does satisfy all six desirable properties, and also captures important nuances of the problem, e.g., both non-exempt masked bias and non-exempt statistically visible bias when they are present together (revisited in Section IV). Our examples also help us understand the utility and limitations of some existing measures that have some provision for exemptions, as we discuss next.

C. *Understanding Existing Measures of Fairness with Provision for Exemptions*

- **Conditional Statistical Parity:** This definition [55], [57] is equivalent to $\mathrm{I}(Z; \hat{Y} \mid X_c) = 0$. Therefore, it has similar utility and limitations as Candidate Measure 1 ($\mathrm{I}(Z; \hat{Y} \mid X_c)$). It resolves some limitations of both statistical parity and equalized odds. However, it gives a false positive conclusion in detecting non-exempt bias in Counterexample 1 (the example of counterfactually fair college admissions), where there is no causal influence of Z on $\hat{Y}$ but $\mathrm{I}(Z; \hat{Y} \mid X_c) > 0$. Because this is an observational measure, it is not able to distinguish between scenarios where there is causal influence of Z on $\hat{Y}$ (non-exempt masked bias in housing ads; Counterexample 2) and where there is not (Counterexample 1), even if $\mathrm{I}(Z; \hat{Y} \mid X_c) > 0$ in both (elaborated further in relation to our impossibility result in Remark 11 Section V). It also fails to capture non-exempt masked bias when the mask arises from the non-critical features as in Example 6.



(a) Venn diagram representation of overall decomposition

	$I(Z;\hat{Y})$	$I(Z;U_X\|\hat{Y})$	$\mathrm{I}(Z;(\hat{Y},U_X))$
M_{NE}	$M_{V,NE} = Uni(Z{:}\hat{Y}\|X_c)$	$M_{M,NE} = M_{NE} - M_{V,NE}$	
M_E	$M_{V,E} = Red(Z{:}(\hat{Y},X_c))$	$M_{M,E} = M_E - M_{V,E}$	
$\mathrm{I}(Z;(\hat{Y},U_X))$			

(b) Tabular representation of overall decomposition

Fig. 7: Overall decomposition of total bias $\mathrm{I}(Z;(\hat{Y},X_c))$ into four non-negative components, namely, non-exempt visible bias $M_{V,NE}$, exempt visible bias $M_{V,E}$, non-exempt masked bias $M_{M,NE}$ and exempt masked bias $M_{M,E}$.

- **Path-Specific Counterfactual Fairness:** Path-specific counterfactual fairness [21] is a purely causal notion of fairness which exempts the causal influence of Z along selected paths. Based on this idea, we proposed Candidate Measure 3 in Section III-B. However, Counterexample 3 (the example of discrimination by unmasking) captures some of its limitations, when there is synergistic or joint information about Z present in X_c and X_g that appears in $\hat{Y}$ that cannot be attributed to any one of them alone. Furthermore, sometimes the influence of Z can cancel along two paths so that the final output has no influence of Z, e.g., the example in Remark 7. For such scenarios, this measure alone can lead to false positive conclusions about non-exempt bias, and might need to be used in conjunction with a measure of total bias (e.g., $\mathrm{CCI}(Z \to \hat{Y})$).
- **Justifiable Fairness:** A model is said to be justifiably fair [56] if $\mathrm{I}(Z;\hat{Y} \mid X_s) = 0$ for all sets $X_s \subseteq X$ such that $X_c \subseteq X_s$. This measure addresses several concerns of the previously stated measures, including capturing several forms of non-exempt masked bias. However, it also gives false positive conclusion in Counterexample 1 (the example of counterfactually fair college admissions), which shows no causal influence of Z on $\hat{Y}$ but $\mathrm{I}(Z;\hat{Y} \mid X_c) > 0$. Because this is an observational measure, it is not able to distinguish between scenarios where there is causal influence of Z on $\hat{Y}$ and where there is not, even if $\mathrm{I}(Z;\hat{Y} \mid X_c) > 0$ in both (elaborated further in relation to our impossibility result in Remark 11 Section V).
  Another limitation of such an individual feature-based conditioning arises when the causal effects of both Z and an independent latent factor are present in the same feature, e.g., different digits of a zip-code, and it is not known in advance whether to condition on the entire zip-code or its sub-portions like the individual digits.

  Example 7. *Let $X_g = [Z, U_{X_1}]$ be a single multivariate feature, e.g., two bits of a number and $X_c = \phi$, and the output be $\hat{Y} = Z \oplus U_{X_1}$ where Z and U_{X_1} are i.i.d. Bern(½).*

  In this example, as long as one treats X_g as a single feature, the model will be deemed *justifiably fair* because $\mathrm{I}(Z;\hat{Y} \mid X_g) = 0$ and $\mathrm{I}(Z;\hat{Y}) = 0$. But, this is a case of non-exempt masked bias. It is necessary to have an advance suspicion of this possible nature of the true SCM to be able to condition on the two bits of X_g separately. This definition captures the non-exempt masked bias in this example if the sub-portions of any single feature are defined in advance.

IV. Understanding the Overall Decomposition

In this section, we demonstrate how our proposed quantification enables a *non-negative* information-theoretic decomposition of the total bias $\mathrm{I}(Z;(\hat{Y},U_X))$ into four components, that can be interpreted as: statistically visible non-exempt bias, statistically visible exempt bias, masked non-exempt bias and masked exempt bias (also see Fig. 7).

Theorem 2 (Non-negative Decomposition of Total Bias)**.** *The total bias can be decomposed into four components as follows:*

$$\mathrm{I}(Z;(\hat{Y},U_X)) = M_{V,NE} + M_{V,E} + M_{M,NE} + M_{M,E}. \tag{11}$$

Here $M_{V,NE} = \mathrm{Uni}(Z:\hat{Y}|X_c)$ and $M_{V,E} = \mathrm{Red}(Z:(\hat{Y},X_c))$. These two terms add to form $\mathrm{I}(Z;\hat{Y})$ which is the total statistically visible bias. Next, $M_{M,NE} = M^*_{NE} - M_{V,NE}$ where M^*_{NE} is our proposed measure of non-exempt bias (Definition 7), and $M_{M,E} = \mathrm{I}(Z;\hat{Y},U_X) - \mathrm{I}(Z;\hat{Y}) - M_{M,NE}$. All of these components are non-negative.

The decomposition of total bias into a summation of these four terms is trivial. What remains to be shown is that these four terms are non-negative (details provided in Appendix D-A).

Interpretation of the four components:

Here $M_{V,NE} = \mathrm{Uni}(Z:\hat{Y}|X_c)$ can be interpreted as the non-exempt statistically visible bias (as also motivated in Section III-B). The remaining part of the statistically visible bias (recall Definition 5), i.e., $\mathrm{I}(Z;\hat{Y}) - \mathrm{Uni}(Z:\hat{Y}|X_c) = \mathrm{Red}(Z:(\hat{Y},X_c))$ then becomes the exempt statistically visible bias ($M_{V,E}$). This also agrees with the intuition that the redundant information about Z statistically visible in both $\hat{Y}$ and Z represents the exempt statistically visible bias.

Now that we have a measure of non-exempt bias (M^*_{NE}) and a measure of non-exempt statistically visible bias ($M_{V,NE}$), we can interpret their difference as the non-exempt masked bias, i.e., $M_{M,NE} = M^*_{NE} - M_{V,NE} = M^*_{NE} - \mathrm{Uni}(Z : \hat{Y}|X_c)$. It also agrees with the intuition that non-exempt masked bias is the part of non-exempt bias that $\mathrm{Uni}(Z : \hat{Y}|X_c)$ alone fails to capture. For instance, recall Counterexample 2 where $\hat{Y} = Z \oplus U_{X_1}$ and $X_c = U_{X_1}$. Here, $\mathrm{I}(Z; \hat{Y}) = 0$, implying $M_{V,NE} = \mathrm{Uni}(Z : \hat{Y}|X_c) = 0$. But, $M^*_{NE} = 1$ bit (supporting derivation in Appendix C-A; see the proof of Theorem 1 under Property 6). Therefore, the non-exempt masked bias $M_{M,NE} = M^*_{NE} - M_{V,NE} = 1$ bit here, which is in agreement with our intuition of non-exempt masked bias. Lastly, the remaining component $M_{M,E} = \mathrm{I}(Z; \hat{Y}, U_X) - \mathrm{I}(Z; \hat{Y}) - M_{M,NE}$ is interpreted as the exempt masked bias. For instance, recall Counterexample 4 where $\hat{Y} = X_c = Z + U_{X_1}$ with $Z, U_{X_1} \sim i.i.d.$ Bern(½). Here, the total bias $\mathrm{I}(Z; \hat{Y}, U_X) = 1$ bit, but the statistically visible bias $\mathrm{I}(Z; \hat{Y}) = 0.5$ bits which means that there is masked bias present. Our intuition is that this masked bias should be entirely exempt because there is no non-exempt bias in this example. This is in agreement with the value that we obtain, i.e., $M_{M,E} = \mathrm{I}(Z; \hat{Y}, U_X) - \mathrm{I}(Z; \hat{Y}) - M_{M,NE} = 0.5$ bits. This is because $M_{M,NE}$ and $M_{V,NE}$ are both non-negative sub-components of M^*_{NE}, and $M^*_{NE} = 0$ (from the Markov chain $(Z, U_{X_1}, U_{X_2}) - X_c - \hat{Y})$).

Remark 9 (On conditioning to capture masked bias)**.** *Conditioning on a random variable G leading to $\mathrm{I}(Z; \hat{Y} \mid G) > I(Z; \hat{Y})$ can sometimes detect masked bias, if conditioning exposes more bias than what was already visible. For example, $I(Z; \hat{Y} \mid X_c)$ can detect masked bias if the mask is of the form $g(X_c)$, e.g., in Counterexample 2 (a special case of the canonical example of masking with $X_c = U_{X_1}$ and $\hat{Y} = Z \oplus U_{X_1}$). However, conditioning on any random variable G leading to $\mathrm{I}(Z; \hat{Y} \mid G) > I(Z; \hat{Y})$ cannot always be interpreted as a case of masked bias because this can sometimes lead to a false positive conclusion in detecting masked bias, e.g., in Counterexample 1 where $\hat{Y} = U_{X_1}$ and $X_c = Z + U_{X_1}$. If G is chosen as X_c, then $\mathrm{I}(Z; \hat{Y} \mid X_c) > I(Z; \hat{Y})$ even though there is no bias here at all (recall $\mathrm{CCI}(Z \to \hat{Y}) = 0$). For completeness, we therefore include another result here (Lemma 3) that clarifies when conditioning can correctly capture masked bias.*

Lemma 3 (Conditioning to Capture Masked Bias)**.** *The following two statements are equivalent:*
- *Masked bias $\mathrm{I}(Z; (\hat{Y}, U_X)) - \mathrm{I}(Z; \hat{Y}) > 0$.*
- *$\exists$ a random variable G of the form $G = g(U_X)$ such that $\mathrm{I}(Z; \hat{Y} \mid G) - \mathrm{I}(Z; \hat{Y}) > 0$.*

Without knowledge of the true causal model, such a $G = g(U_X)$ may be difficult to determine from observational data alone, because the observational data can be a function of both Z and U_X. This serves as the motivation behind our impossibility result on observational measures, that we state next.

V. Impossibility Result

Theorem 3 (Impossibility of Observational Measures)**.** *No observational measure of non-exempt bias simultaneously satisfies all six desirable properties.*

Proof of Theorem 3. Observe the two examples here:

Example 8 (A Case of No Bias)**.** *Let $X_c = Z \oplus U_{X_1}$, $X_g = Z$ and $\hat{Y} = X_c \oplus X_g = U_{X_1}$ where Z and U_{X_1} are both independent and identically distributed as Bern(½).*

Example 9 (A Case of Non-Exempt Bias)**.** *Let $X_c = U_{X_1}$, $X_g = Z$ and $\hat{Y} = X_c \oplus X_g = Z \oplus U_{X_1}$ where Z and U_{X_1} are both independent and identically distributed as Bern(½).*

In Example 8, the influences of Z cancel each other and there is no total bias. So, the non-exempt bias should be zero by Property 4 (Zero Influence). However, Example 9 is the canonical example of non-exempt masked bias where there is non-exempt bias present, and hence the non-exempt bias should be non-zero by Property 6 (Non-Exempt Masked Bias). But, for both of these examples, the joint distribution of the observables $(Z, X_c, X_g, \hat{Y})$ is the same which means that no observational measure can distinguish between these two cases. This proves the result. ☐

Remark 10 (Alternative Examples)**.** *In fact, we can show that no observational measure can satisfy Property 6. Consider a scenario of no bias given by: $X_c = \phi$, $X_g = (Z \oplus U_{X_1}, Z)$ and $\hat{Y} = U_{X_1}$. For this example, the Markov chain $Z - X_c - (\hat{Y}, U_{X_1})$ holds implying that $M_{NE} = 0$ by Property 6. Alternatively, consider a scenario of non-exempt bias given by: $X_c = \phi$, $X_g = (U_{X_1}, Z)$ and $\hat{Y} = Z \oplus U_{X_1}$ which is again a variant of the canonical example of non-exempt masked discrimination. Let Z and U_{X_1} be independent and identically distributed as Bern(½). Then, no purely observational measure can distinguish between these two scenarios because $(Z, X_c, X_g, \hat{Y})$ have the same joint distribution.*

Remark 11 (Revisiting Conditional Statistical Parity and Justifiable Fairness)**.** *For both Examples 8 and 9, we observe that conditional mutual information $\mathrm{I}(Z; \hat{Y} \mid X_c) > 0$. Because $\mathrm{I}(Z; \hat{Y} \mid X_c)$ is an observational measure, it fails to distinguish between whether there is causal influence of Z or not in $\hat{Y}$. Existing observational definitions of fairness, e.g., conditional statistical parity and justifiable fairness would also not be able to distinguish between these two examples. One needs counterfactual measures to be able to distinguish between them, such as the counterfactual measure proposed in this work.*

Nevertheless, because counterfactual measures are difficult to realize in practice, we examine the following observational measures of non-exempt bias that satisfy only a few of Properties 1-6.

VI. Observational Relaxations of our Proposed Counterfactual Measure: Utility and Limitations

In this section, we propose three observational measures of non-exempt bias and discuss their utility and limitations.

Observational Measure 1. $M_{NE} = \mathrm{Uni}(Z : \hat{Y}|X_c)$.

Utility: This measure satisfies several desirable properties as stated here:

Lemma 4. *[Fairness Properties of* $\mathrm{Uni}(Z : \hat{Y}|X_c)$*] The measure* $\mathrm{Uni}(Z : \hat{Y}|X_c)$ *satisfies Properties 2, 3, 4, and 5.*

The proof is in Appendix E. Importantly, note that, $\mathrm{Uni}(Z : \hat{Y}|X_c)$ satisfies Property 4 which $\mathrm{I}(Z; \hat{Y} \mid X_c)$ does not (recall Counterexample 1). Thus, $\mathrm{Uni}(Z : \hat{Y}|X_c)$ does not give false positive conclusions in detecting non-exempt bias if a model is counterfactually fair.

This measure may be preferred over our other observational measures when one wants to prioritize avoiding false positive quantification of non-exempt bias when a model is counterfactually fair. Recall that, $\mathrm{Uni}(Z : \hat{Y}|X_c)$ is a measure of non-exempt, statistically visible bias. *It correctly captures the entire non-exempt bias when non-exempt masked bias is absent.*

Limitations: It does not quantify any non-exempt masked bias (Property 6). This is because $\mathrm{Uni}(Z : \hat{Y}|X_c)$ is a sub-component of the statistically visible bias $\mathrm{I}(Z; \hat{Y})$, and hence always goes to 0 whenever the statistically visible bias $\mathrm{I}(Z; \hat{Y}) = 0$ (recall Example 3). It also does not satisfy Property 1 because when $X_c = \phi$, we have $\mathrm{Uni}(Z : \hat{Y}|X_c) = \mathrm{I}(Z; \hat{Y})$, which is only the statistically visible bias but not the total bias in a counterfactual sense (i.e., $\mathrm{I}(Z; \hat{Y}, U_X)$).

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

Utility: This measure also satisfies several desirable properties, as stated here:

Lemma 5. *[Fairness Properties of* $\mathrm{I}(Z; \hat{Y} \mid X_c)$*] The measure* $\mathrm{I}(Z; \hat{Y} \mid X_c)$ *satisfies Properties 3 and 5.*

The proof is in Appendix E. We note that, while it does not satisfy Property 6 in its entirely, it does capture some scenarios of non-exempt masked bias. E.g., it can detect the non-exempt masked bias in Counterexample 2 which $\mathrm{Uni}(Z : \hat{Y}|X_c)$ is not able to, even though they both fail to detect the non-exempt masked bias in Example 6. In general, $\mathrm{I}(Z; \hat{Y} \mid X_c)$ can detect non-exempt masked bias when the "mask" is entirely derived from the critical features, i.e., $G = g(X_c)$.

Limitations:

It can sometimes lead to false positive conclusion about non-exempt bias, e.g., in Counterexample 1 (does not satisfy Property 4). It also does not satisfy Property 2 because clearly $\mathrm{I}(Z; \hat{Y} \mid X_c)$ may be greater or less that $\mathrm{I}(Z; \hat{Y})$ (recall Counterexample 2). It also does not satisfy Property 1 because when $X_c = \phi$, we have $\mathrm{I}(Z; \hat{Y} \mid X_c) = \mathrm{I}(Z; \hat{Y})$, which is only the statistically visible bias but not the total bias in a counterfactual sense (i.e., $\mathrm{I}(Z; \hat{Y}, U_X)$).

Observational Measure 3. $M_{NE} = \mathrm{I}(Z; \hat{Y} \mid X_c, X')$ where X' consists of certain features in X_g.

Utility and Limitations: This is somewhat of a heuristic relaxation that only satisfies Property 3. However, while it does not satisfy any of the other properties in their entirety, it can still lead to the desirable quantification in several examples where the previous two measures may not be successful if X' is chosen appropriately. For example, recall Example 6 where $\hat{Y} = Z \oplus U_{X_1}$ with $X_g = (Z, U_{X_1})$. With some partial knowledge or assumption about the SCM, if we choose $X' = U_{X_1}$, then $\mathrm{I}(Z; \hat{Y} \mid X_c, X') > 0$ for this example even though $\mathrm{I}(Z; \hat{Y} \mid X_c) = 0$. Thus, this measure is able to detect some more scenarios of non-exempt masked bias that $\mathrm{I}(Z; \hat{Y} \mid X_c)$ cannot, i.e., when the mask is of the form $G = g(X_c, X')$. It can also sometimes avoid false positive quantification of non-exempt bias if X' is chosen appropriately, e.g., in Counterexample 1 if $X' = U_{X_1}$. Thus, under partial knowledge or assumption about the true SCM, this measure can correctly capture the non-exempt bias in many scenarios where the previous two measures may not be successful.

Lastly, one may also consider using various combinations of these measures, e.g., $\mathrm{Uni}(Z : \hat{Y}|X_c) + \mathrm{I}(Z; \hat{Y} \mid X')$, or $\mathrm{I}(Z; \hat{Y} \mid X_c) + \mathrm{I}(Z; \hat{Y} \mid X')$, or $\mathrm{Uni}(Z : \hat{Y}|X_c) + \mathrm{Syn}(Z : (\hat{Y}, X'))$, that can also approximate our proposed measure in several scenarios if X' is chosen appropriately based on partial knowledge or assumptions about the true SCM.

VII. Case Studies Demonstrating Practical Application in Training

Here, we discuss some case studies on both simulated and real data. As a first step in this direction, we will only use one of our proposed measures, namely, $\mathrm{I}(Z; \hat{Y} \mid X_c)$, and choose a simple correlation-based estimate for it (elaborated further in Section VII-A) in this work. In future work, we will explore some extensions of these ideas to estimating unique information, as well as, some alternate methods of estimation altogether [25], [35], [55], [68], [80], as discussed in Section VIII.

A. Case Study on Simulated Data

Setup: The goal is to decide whether to show ads for an editor job requiring English proficiency, based on whether a score generated from internet activity is above a threshold. $Z \sim Bern(1/2)$ is a protected attribute denoting whether a person is a native English speaker or not. Now, consider three features $X = (X_1, X_2, X_3)$, where $X_c = X_1$ and $X_g = (X_2, X_3)$: (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

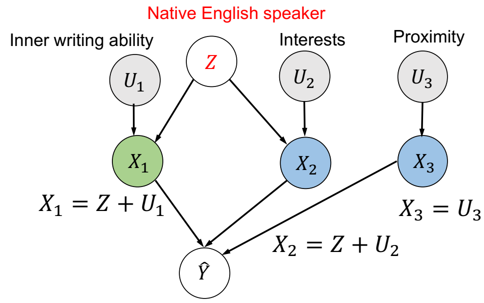


Fig. 8: Structural causal model corresponding to the simulated case study: The goal is to decide whether to show ads for an editor job requiring English proficiency, based on three features using logistic regression. The three features are (i) X_1: a score based on online writing samples; (ii) X_2: a 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 proximity. $Z \sim Bern(1/2)$ is a protected attribute denoting whether a person is a native English speaker or not.

compared to websites of other languages; and (iii) X_3: a preference score based on geographical proximity. Suppose the true SCM is as follows (see Fig. 8): $X_1 = Z + U_1$, $X_2 = Z + U_2$, and $X_3 = U_3$ 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 given by $Y = \mathbb{1}(S \geq 1)$ indicating whether $S \geq 1$ or not.

Models trained: We will train a logistic regression model of the form $\hat{Y} = 1/(1 + e^{-(w^T X + b)})$. The model produces an output value $\hat{Y}$ between 0 and 1. Using this model, one 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 (No Fairness):

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

Loss L_2 (Mutual Information (MI) Regularizer):

$$\min_{w,b} L_{\text{Cross Entropy}}(Y, \hat{Y}) + \lambda \widetilde{\mathrm{I}}(Z; \hat{Y}),$$

where (i) λ is the regularization constant; and (ii) $\widetilde{\mathrm{I}}(Z; \hat{Y}) = -\frac{1}{2} \log(1 - \rho^2_{Z,\hat{Y}})$ 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 [79].

Loss L_3 (Conditional Mutual Information (CMI) Regularizer):

$$\min_{w,b} L_{\text{Cross Entropy}}(Y, \hat{Y}) + \lambda \widetilde{\mathrm{I}}(Z; \hat{Y} \mid X_c),$$

where again (i) λ is the regularization constant; and (ii) $\widetilde{\mathrm{I}}(Z; \hat{Y} \mid X_c)$ is given by:

$$\widetilde{\mathrm{I}}(Z; \hat{Y} \mid X_c) = \sum_{i=1}^{n} \Pr(X_c \in \text{Bin } i) \widetilde{\mathrm{I}}(Z; \hat{Y} \mid X_c \in \text{Bin } i) = -\frac{1}{2} \sum_{i=1}^{n} \Pr(X_c \in \text{Bin } i) \log(1 - \rho^2_{Z,\hat{Y},i}),$$

where the range of X_c is divided into n discrete bins, and $\rho_{Z,\hat{Y},i}$ is the conditional correlation of $\hat{Y}$ and Z given X_c is in the i-th discrete bin.

Observations (Fig. 9 and Table III): We plot the values of accuracy and non-exempt bias (CMI) in Fig. 9 for the three different loss functions, by varying the regularization constant λ, wherever applicable. While we only include few values of λ in Table III for brevity, the plot in Fig. 9 is based on all integer values of λ ranging from 1 to 10 (100 simulations of 7000 iterations each with batch size 200).

Intuition behind the results (Fig. 10): To understand the results better, we specifically look into the histogram of the predictions for all candidates and the histogram for only those candidates with similar value of the critical feature (e.g., $X_c \geq 0.5$). For Loss 1, the model learns to place equal weight on all three features, consistent with the historic scores. Thus, it attains a high accuracy. But, because the historic scores are correlated with browsing history (X_2), even when a non-native speaker has good writing score, they are not be shown an ad due to their browsing history (hence, the high non-exempt bias). Loss 2 (MI Regularizer) 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 III). However, Loss 3 (CMI Regularizer) is able to reduce the importance (weight) of browsing history relative to online writing scores, leading to an intermediate accuracy between Loss 1 and 2 for same non-exempt bias (also see Fig. 9). In a sense, our measure enables individuals with similar X_c to be treated similarly.

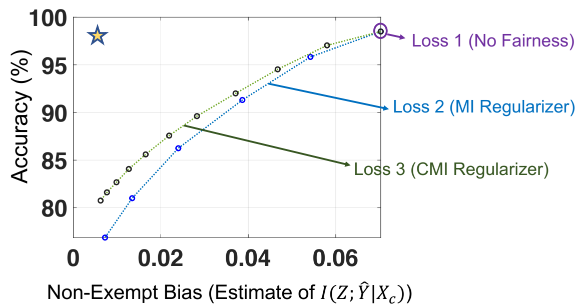


Fig. 9: Trade-off between accuracy and non-exempt bias (CMI): For Loss 1 (No Fairness), the trained model achieves high accuracy but also has high non-exempt bias (CMI). For both Loss 2 (MI Regularizer) and Loss 3 (CMI Regularizer), the non-exempt bias reduces, but so does accuracy. But, Loss 3 achieves a much better trade-off between accuracy and non-exempt bias than Loss 2. For instance, Loss 3 (with $\lambda = 7$) achieves 84.07% accuracy while Loss 2 (with $\lambda = 4$) achieves only 81.00% while attaining the same value of non-exempt bias (CMI). λ takes all integer values from 1 to 10.

TABLE III: Observations after training a logistic regression model ($\hat{Y} = 1/(1 + e^{-(w^T X + b)})$) using three loss functions with different fairness criteria (100 simulations of 7000 iterations each with batch size 200).

Setup	λ	Accuracy (SD.)	CMI (SD.)	$-\frac{w_1}{b}$ (SD.)	$-\frac{w_2}{b}$ (SD.)	$-\frac{w_3}{b}$ (SD.)
Loss L_1	N.A.	98.46 (0.10)	0.0703 (0.004)	1.083 (0.003)	1.083 (0.003)	1.075 (0.003)
Loss L_2	$\lambda = 1$	95.82 (0.27)	0.0542 (0.004)	1.108 (0.006)	1.108 (0.006)	1.449 (0.018)
	$\lambda = 2$	91.31 (0.38)	0.0386 (0.003)	1.118 (0.01)	1.119 (0.01)	1.954 (0.052)
	$\lambda = 5$	76.87 (0.005)	0.0072 (0.0012)	1.038 (0.0311)	1.034 (0.0312)	5.183 (0.1441)
	$\lambda = 10$	70.17 (0.67)	0.0011 (0.0005)	0.9957 (0.1372)	0.9990 (0.1518)	13.9077 (0.9882)
Loss L_3	$\lambda = 1$	97.04 (0.20)	0.0580 (0.004)	1.154 (0.005)	1.037 (0.003)	1.294 (0.009)
	$\lambda = 2$	94.52 (0.29)	0.0467 (0.003)	1.243 (0.0109)	0.962 (0.004)	1.511 (0.0180)
	$\lambda = 5$	87.57 (0.34)	0.0219 (0.002)	1.563 (0.0195)	0.602 (0.0145)	2.0685 (0.0287)
	$\lambda = 10$	80.76 (0.50)	0.0062 (0.0011)	2.0495 (0.0272)	0.0235 (0.0248)	2.5695 (0.0339)

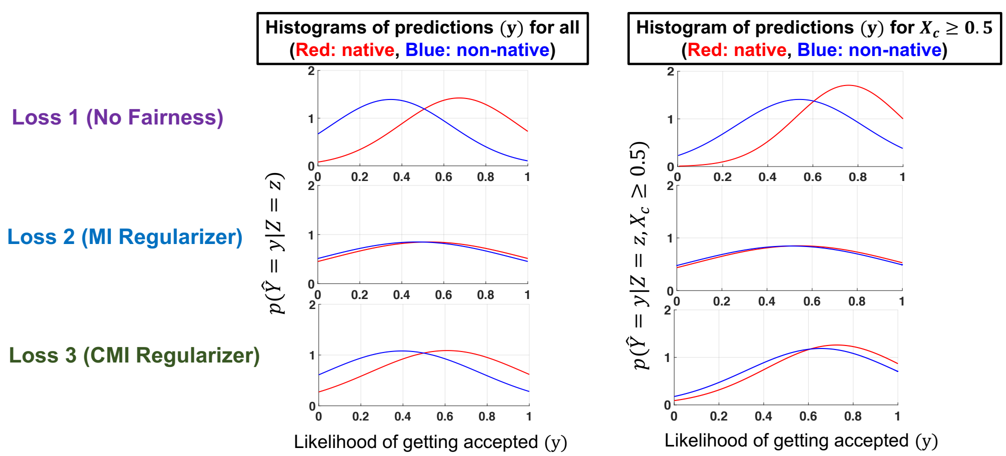


Fig. 10: Histogram of model outputs for a particular value of $\lambda = 10$: (Left) $p(\hat{Y} = y \mid Z = z)$ for $z = 0, 1$; (Right) $p(\hat{Y} = y \mid X_c \geq 0.5, Z = z)$ for $z = 0, 1$. Loss 2 (MI Regularizer) brings $p(\hat{Y} = y \mid Z = z)$ closer for Z=0 and 1 by placing higher weight on a less important feature (proximity score). But this distorts the histograms significantly and reduces the accuracy (see Table III). Loss 3 (CMI Regularizer) still retains some biases in the overall histogram ($p(\hat{Y} = y \mid Z = z)$), and only brings $p(\hat{Y} = y|X_c{\geq}0.5, Z = z)$ approach each other for $z = 0$ and 1, so that candidates with similar critical feature X_c are treated similarly and not discriminated any further based on Z.

B. Case Study on Real Data: `Adult` *Dataset*

Setup: The Adult dataset [81], also known as the Census income dataset, consists of 14 features such as age, educational qualification, etc., and the prediction goal is to predict whether the income is greater than 50K. This dataset is widely used in existing fairness literature (see [26]), because such data might be representative of the data used in highly consequential applications, such as, lending, showing expensive ads, etc. In this work, we choose gender as the protected attribute (Z) in the

Adult dataset. Our set of input features (X) consist of all the other features except gender, and our critical feature (X_c) is hours-per-week.

Models trained: We train a deep neural network (multi-layer perceptron) on this dataset. The model takes all the features except gender as input (with one hot encoding of all categorical variables). The input layer is followed by three hidden layers, each having 32 neurons with ReLu activation and dropout probability 0.2. Finally, the output layer consists of a single neuron with sigmoid activation that produces an output value between 0 and 1 (denoting likelihood of acceptance). We again define the three loss functions as before: (i) only binary cross-entropy loss with no fairness criterion; (ii) binary cross-entropy loss with Mutual Information (MI) regularizer; and (iii) binary cross-entropy loss with Conditional Mutual Information (CMI) regularizer.

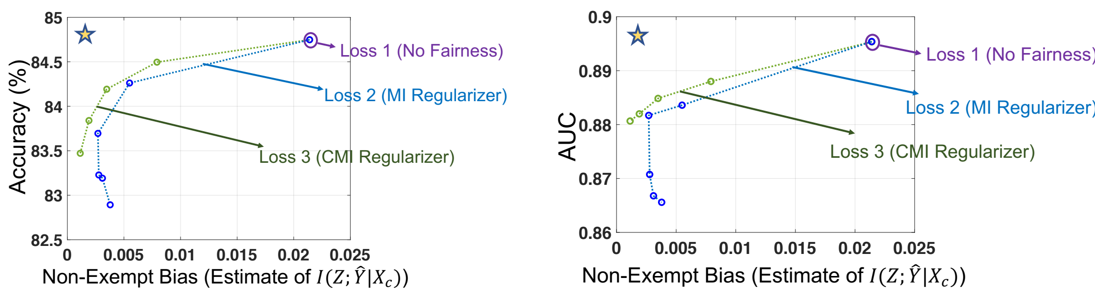


Fig. 11: Trade-off between accuracy/AUC and non-exempt bias (CMI): For Loss 1 (No Fairness), the trained model achieves high accuracy but also has high non-exempt bias (CMI). For both Loss 2 (MI Regularizer) and Loss 3 (CMI Regularizer), the non-exempt bias reduces, but so does accuracy. But, Loss 3 achieves a much better trade-off between accuracy and non-exempt bias than Loss 2. Interestingly, we note that, as we increase the value of the regularization constant (λ), the non-exempt bias (CMI) for Loss 2 begins to increase after a point. This is because we enter into a "high-synergy" regime where CMI is higher than MI, and MI still decreases but without decreasing CMI (more intuition in Fig. 12).

Observations (Fig. 11): Similar to the previous case study, we plot the values of accuracy/AUC and non-exempt bias (CMI) in Fig. 11 for the three different loss functions by varying the regularization constant (λ) wherever applicable. The plot is based on $\lambda = 1, 2, 4, 6$ (averaged over 40 simulations of 300 epochs each with batch-size 1000).

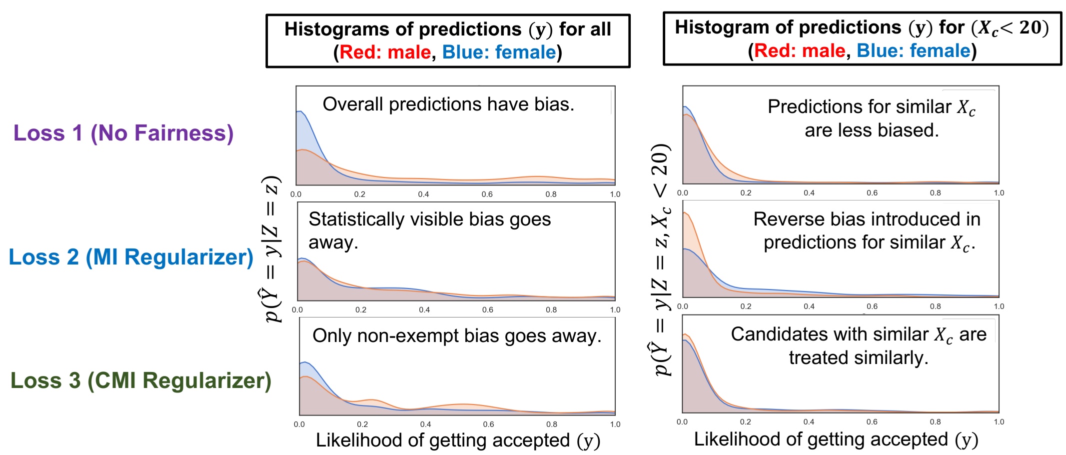


Fig. 12: Histogram of model outputs for a particular value of $\lambda = 4$: (Left) $p(\hat{Y} = y \mid Z = z)$ for $z = 0, 1$; (Right) $p(\hat{Y} = y \mid X_c < 20, Z = z)$ for $z = 0, 1$. Loss 2 (MI Regularizer) brings $p(\hat{Y} = y \mid Z = z)$ closer for Z=0 and 1. But this introduces some reverse bias among candidates with similar value of critical feature. Here CMI>MI, and non-exempt bias (CMI) persists even though MI is significantly reduced. Loss 3 (CMI Regularizer) still retains some biases in the overall histogram ($p(\hat{Y} = y \mid Z = z)$), but it brings $p(\hat{Y} = y|X_c{<}20, Z = z)$ approach each other for $z = 0$ and 1, so that candidates with similar critical feature X_c are treated similarly and not discriminated any further based on Z.

Intuition behind results (Fig. 12): To understand the results better, we again specifically look into the histogram of the predictions for all candidates and the histogram for only those candidates with similar value of the critical feature (e.g., $X_c < 20$). For Loss 1, the overall predictions have bias, but predictions for similar X_c are less biased (i.e., MI > CMI). Using Loss 2 (MI Regularizer), the overall statistically visible bias goes away, but when the histogram of candidates with similar value of critical feature are examined specifically, we notice a "reverse" bias introduced. Thus, CMI is quite higher than MI, i.e., non-exempt bias (CMI) persists even though MI is significantly reduced (statistical parity is almost attained). However, Loss 3 (CMI Regularizer) is able to selectively reduce the non-exempt bias in the overall predictions, while attempting to treat candidates with similar X_c similarly irrespective of their gender.

VIII. Discussion and Conclusion

On Choice of Critical Features and Connections with Explainability: In this work, as also in some existing works on fairness [19], [44], we assume that the critical features are known. We adopt a viewpoint stated in [48] which suggests that "We can't just rely on the math; we still need a human person applying human judgements." Since most of these exemptions are embedded in law and social science [43], [45]–[47], we believe that fairness researchers need to collaborate with social scientists and lawyers in order to determine which set of features can be designated as critical for a particular application.

This work also shares close connections with the field of *explainability* in machine learning [17], [64], [82], and motivates several related research problems, e.g., how to check or explain if certain features contributed to the bias in a model, or how to incorporate exemptions in applications, such as, image processing, where certain neurons in an intermediate hidden layer might need to be exempted instead of the input layer because they often have more interpretability [82].

On Better Understanding of Observational Measures: Our proposed counterfactual measure and the desirable properties help in evaluation of observational measures in practice, and understand their utility and limitation, i.e., what they capture and miss. Finally, in applications where when the true SCM is known or can be evaluated from the data [50, Chapters 4,7], the proposed measure exactly captures the non-exempt bias.

On Uniqueness, Operational Meaning and Further Generalizations: We acknowledge that we do not prove uniqueness of our measure with respect to the desirable properties, and neither do we show that the properties are exhaustive (recall Remark 5 in Section III-B). This is an interesting direction of future work. However, there may also be value in the fact that the properties do not yield a unique measure: this allows for tuning the measure based on the application. E.g., Shannon established uniqueness on entropy with respect to **some** properties in [74] but subsequent applications have still led to the use of modified measures, e.g. Renyi entropy [25], [68], [75], [76].

Deriving the exact operational meaning of our proposed counterfactual measure is also an interesting direction of future work. Nonetheless, the proposed measure does satisfy our stated desirable properties and capture important aspects of the problem, e.g., statistically visible and masked biases. Furthermore, our measure can also be modified to account for further functional generalizations.

First notice, that our proposed Property 6 is a special case of the following statement:

If $(Z, f_a(U_X))$–X_c–$(\hat{Y}, f_b(U_X))$ form a Markov chain for any deterministic functions $f_a(\cdot)$ and $f_b(\cdot)$ such that $f_a(U_X) \perp\!\!\!\perp f_b(U_X)$ and $\mathrm{H}(U_X){=}\mathrm{H}(f_a(U_X)) + \mathrm{H}(f_b(U_X))$, then $M_{NE}{=}0$.

To account for this more general property, our proposed measure might be modified as follows:

$$\min_{f_a(U_X), f_b(U_X)} \mathrm{Uni}((Z, f_a(U_X)) : (\hat{Y}, f_b(U_X)) | X_c), \tag{12}$$

such that $f_a(U_X) \perp\!\!\!\perp f_b(U_X)$ and $\mathrm{H}(U_X) = \mathrm{H}(f_a(U_X)) + \mathrm{H}(f_b(U_X))$. This measure also satisfies all the other desirable properties. In this work, we restrict ourselves to $f_a(U_X)$ and $f_b(U_X)$ being disjoint subsets of U_X for simplicity, computability and ease of understanding. Future work will explore how different assumptions on the SCM restrict the class of f_a and f_b.

On Understanding Other Forms of Masked Bias: Let us revisit the discussion from Section III-B that not all forms of masked discrimination are necessarily undesirable. E.g., if U_{X_1} is a random coin flip in Example 3, then performing $\hat{Y} = Z \oplus U_{X_1}$ randomizes the race, and can even be regarded as a preventive measure against discrimination. However, keeping the mathematics of the example same, if U_{X_1} instead denotes whether one's income is above a threshold, then the model is unfair. It is an interesting future direction to examine how to quantify non-exempt discrimination while allowing the user with more flexibility on what latent factors are allowed to mask Z.

On Estimation of Mutual Information, Conditional Mutual Information and Unique Information: In general, it is difficult to directly incorporate these information-theoretic measures as a regularizer with the loss function (see [80], [83] and the references therein). In this work, as a first step, we used a simple correlation-based estimate for mutual and conditional mutual information under a Gaussian assumption. We believe that similar ideas can be extended to unique information as well under Gaussian assumptions, building on [84]. Examining alternate methods of incorporating our proposed measures as regularizer (using or building upon techniques proposed in [25], [35], [55], [68], [80]) is an interesting direction of future work.

Acknowledgment

This work was supported by an NSF Career Award and NSF grant CNS-1704845 as well as by DARPA and the Air Force Research Laboratory under agreement number FA8750-15-2-0277. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of DARPA, the Air Force Research Laboratory, the National Science Foundation, or the U.S. Government. S. Dutta was supported by the K&L Gates Presidential Fellowship in Ethics and Computational Technologies 2019 and the Axel Berny Graduate Fellowship 2019. P. Venkatesh was supported by a Fellowship in Digital Health from the Center for Machine Learning and Health at Carnegie Mellon University.

References

[1] S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover, "An information-theoretic quantification of discrimination with exempt features," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
[2] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 2012, pp. 214–226.
[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," *arXiv preprint arXiv:1803.02453*, 2018.
[4] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.
[5] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.
[6] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 107–118.
[7] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 35–50.
[8] J. Komiyama and H. Shimao, "Two-stage algorithm for fairness-aware machine learning," *arXiv preprint arXiv:1710.04924*, 2017.
[9] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Advances in Neural Information Processing Systems*, 2018, pp. 2791–2801.
[10] A. Ghassami, S. Khodadadian, and N. Kiyavash, "Fairness in supervised learning: An information theoretic approach," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 176–180.
[11] H. Heidari, C. Ferrari, K. Gummadi, and A. Krause, "Fairness behind a veil of ignorance: A welfare analysis for automated decision making," in *Advances in Neural Information Processing Systems*, 2018, pp. 1265–1276.
[12] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," *arXiv preprint arXiv:1507.05259*, 2015.
[13] Z. Lipton, J. McAuley, and A. Chouldechova, "Does mitigating ML's impact disparity require treatment disparity?" in *Advances in Neural Information Processing Systems*, 2018, pp. 8125–8135.
[14] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney, "Data Augmentation for Discrimination Prevention and Bias Disambiguation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 358–364.
[15] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. R. Varshney, "An information-theoretic perspective on the relationship between fairness and accuracy," *arXiv preprint arXiv:1910.07870*, 2019.
[16] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 259–268.
[17] K. R. Varshney, "Trustworthy machine learning and artificial intelligence," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 25, no. 3, pp. 26–29, 2019.
[18] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.
[19] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 656–666.
[20] C. Russell, M. J. Kusner, J. Loftus, and R. Silva, "When worlds collide: Integrating different counterfactual assumptions in fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 6414–6423.
[21] S. Chiappa, "Path-specific counterfactual fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7801–7808.
[22] A. Datta, M. Fredrikson, G. Ko, P. Mardziel, and S. Sen, "Use privacy in data-driven systems: Theory and experiments with machine learnt programs," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 1193–1210.
[23] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018.
[24] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
[25] J. Liao, C. Huang, P. Kairouz, and L. Sankar, "Learning generative adversarial representations (gap) under fairness and censoring constraints," *arXiv preprint arXiv:1910.00411*, 2019.
[26] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.
[27] S. Yeom, A. Datta, and M. Fredrikson, "Hunting for discriminatory proxies in linear regression models," in *Advances in Neural Information Processing Systems*, 2018, pp. 4568–4578.
[28] R. C. Williamson and A. K. Menon, "Fairness risk measures," *arXiv preprint arXiv:1901.08665*, 2019.
[29] D. Wei, K. N. Ramamurthy, and F. d. P. Calmon, "Optimized score transformation for fair classification," *arXiv preprint arXiv:1906.00066*, 2019.
[30] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 2125–2126.
[31] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 560–568.
[32] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar, "A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 2239–2248.
[33] H. Wang, H. Hsu, M. Diaz, and F. P. Calmon, "To split or not to split: The impact of disparate treatment in classification," *arXiv preprint arXiv:2002.04788*, 2020.
[34] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International Conference on Machine Learning*, 2018, pp. 2564–2572.
[35] J. Cho, G. Hwang, and C. Suh, "A fair classifier using mutual information," 2020.
[36] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im) possibility of fairness," *arXiv preprint arXiv:1609.07236*, 2016.
[37] S. Yeom and M. C. Tschantz, "Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews," *arXiv preprint arXiv:1808.08619*, 2018.
[38] M. J. Kusner, C. Russell, J. R. Loftus, and R. Silva, "Causal Interventions for Fairness," *arXiv preprint arXiv:1806.02380*, 2018.
[39] S. Sharma, J. Henderson, and J. Ghosh, "Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models," *arXiv preprint arXiv:1905.07857*, 2019.
[40] C. Dwork, C. Ilvento, and M. Jagadeesan, "Individual fairness in pipelines," *arXiv preprint arXiv:2004.05167*, 2020.
[41] J. H. Hinnefeld, P. Cooman, N. Mammo, and R. Deese, "Evaluating fairness metrics in the presence of dataset bias," *arXiv preprint arXiv:1809.09245*, 2018.
[42] Reuters, "Amazon scraps secret AI recruiting tool that showed bias against women," https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
[43] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[44] F. Kamiran, I. Žliobaitė, and T. Calders, "Quantifying explainable discrimination and removing illegal discrimination in automated decision making," *Knowledge and information systems*, vol. 35, no. 3, pp. 613–644, 2013.
[45] EEOC Website, "US Equal Pay Act," https://www.eeoc.gov/laws/statutes/epa.cfm.
[46] EEOC Website, "Bona-Fide Occupational Qualification (BFOQ)," https://www.eeoc.gov/eeoc/foia/letters/2001/ada_job_related.html.
[47] Dept. of Trade and Industry, *Genuine Occupational Qualifications, A Good Practice Guide for Employers*, 2008.
[48] P. Pandey, "Is your machine learning model biased?" https://towardsdatascience.com/is-your-machine-learning-model-biased-94f9ee176b67.
[49] J. H. Hinnefeld, "Measuring model fairness," https://www.youtube.com/watch?v=V3tmxMf2UH8.
[50] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT press, 2017.
[51] T. Tax, P. Mediano, and M. Shanahan, "The partial information decomposition of generative neural network models," *Entropy*, vol. 19, no. 9, p. 474, 2017.
[52] P. Venkatesh, S. Dutta, and P. Grover, "Information flow in computational systems," *IEEE Transactions on Information Theory (To Appear)*, 2020.
[53] P. Venkatesh, S. Dutta, and P. Grover, "How should we define Information Flow in Neural Circuits?" in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 176–180.
[54] P. Venkatesh, S. Dutta, and P. Grover, "How else can we define Information Flow in Neural Circuits?" in *2020 IEEE International Symposium on Information Theory (ISIT, To Appear)*, 2020.
[55] Anonymous, "Conditional debiasing for neural networks."
[56] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu, "Interventional Fairness: Causal Database Repair for Algorithmic Fairness," in *Proceedings of the 2019 International Conference on Management of Data*, ser. SIGMOD '19. ACM, 2019, pp. 793–810.
[57] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. ACM, 2017, pp. 797–806. [Online]. Available: http://doi.acm.org/10.1145/3097983.3098095
[58] E. W. Weisstein, "n-tuple," https://mathworld.wolfram.com/n-Tuple.html, from MathWorld–A Wolfram Web Resource.
[59] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, 2014.
[60] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.
[61] V. Griffith and C. Koch, "Quantifying synergistic mutual information," in *Guided Self-Organization: Inception*. Springer, 2014, pp. 159–190.
[62] P. K. Banerjee, J. Rauh, and G. Montúfar, "Computing the unique information," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 141–145.
[63] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[64] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 598–617.
[65] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1885–1894.
[66] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," *Knowledge and Information Systems*, vol. 54, no. 1, pp. 95–122, 2018.
[67] A. Henelius, K. Puolamäki, H. Boström, L. Asker, and P. Papapetrou, "A peek into the black box: exploring classifiers by randomization," *Data mining and knowledge discovery*, vol. 28, no. 5-6, pp. 1503–1529, 2014.
[68] J. Liao, L. Sankar, O. Kosut, and F. P. Calmon, "Robustness of maximal α-leakage to side information," *arXiv preprint arXiv:1901.07105*, 2019.
[69] S. Newstead and A. D'Elia, "An investigation into the relationship between vehicle colour and crash risk," *Prevention*, vol. 17, no. 1, pp. 47–56, 2007.
[70] R. Digest, "Here's how different car colors affect your chances of getting into an accident," https://www.rd.com/culture/car-color-accidents-risk/.
[71] S. Garg, M. P. Kim, and O. Reingold, "Tracking and improving information in the service of fairness," in *Proceedings of the ACM Conference on Economics and Computation*, 2019, pp. 809–824.
[72] I. Y. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" in *Advances in Neural Information Processing Systems*, 2018, pp. 3539–3550.
[73] P. K. Banerjee, E. Olbrich, J. Jost, and J. Rauh, "Unique informations and deficiencies," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2018, pp. 32–38.
[74] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
[75] A. Rényi, "On measures of entropy and information," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. The Regents of the University of California, 1961.
[76] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Transactions on Information Theory*, 2019.
[77] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE, 2017, pp. 263–275.
[78] N. Kilbertus, P. Ball, M. Kusner, A. Weller, and R. Silva, "The sensitivity of counterfactual fairness to unmeasured confounding," in *35th Conference on Uncertainty in Artificial Intelligence, UAI 2019*. Association for Uncertainty in Artificial Intelligence, 2019.
[79] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.
[80] S. Mukherjee, H. Asnani, and S. Kannan, "CCMI: Classifier based Conditional Mutual Information estimation," *arXiv preprint arXiv:1906.01824*, 2019.
[81] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml
[82] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in *International Conference on Machine Learning*, 2018, pp. 2668–2677.
[83] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Advances in Neural Information Processing Systems*, 2010, pp. 1849–1857.
[84] A. B. Barrett, "Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems," *Physical Review E*, vol. 91, no. 5, p. 052802, 2015.
[85] J. Rauh, P. Kr. Banerjee, E. Olbrich, and J. Jost, "Unique Information and Secret Key Decompositions," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 3042–3046.

APPENDIX A
COUNTERFACTUAL CAUSAL INFLUENCE (CCI) AND ITS CONNECTION TO COUNTERFACTUAL FAIRNESS

A. Proof of Lemma 1

Here, we first provide a proof of Lemma 1 which shows that our proposed quantification of total bias is zero if and only if $\mathrm{CCI}(Z \to \hat{Y}) = 0$. For ease of reading, we repeat the statement of the lemma here again.

Lemma 1 (Equivalences of CCI)**.** *Consider the aforementioned system model. Let* $\hat{Y} = h(Z, U_X)$ *for some deterministic function* $h(\cdot)$ *and* $Z \perp\!\!\!\perp U_X$*. Then,* $\mathrm{CCI}(Z \to \hat{Y}) = 0$ *if and only if* $\mathrm{I}(Z; (\hat{Y}, U_X)) = 0$*.*

Proof of Lemma 1. From the definition of CCI (Definition 3 in Section II-B),

$$\begin{aligned} &\mathrm{CCI}(Z \to \hat{Y}) = \mathbb{E}_{Z,Z',U_X}\left[|h(Z,U_X) - h(Z',U_X)|\right] \\ &= \sum_{z_1,z_2,u_x} \Pr(Z = z_1, Z' = z_2, U_X = u_x)|h(z_1,u_x) - h(z_2,u_x)| \\ &= \sum_{z_1,z_2,u_x} \Pr(Z = z_1)\Pr(Z' = z_2)\Pr(U_X = u_x)|h(z_1,u_x) - h(z_2,u_x)| \qquad \text{[from independence].} \end{aligned} \tag{13}$$

The summation consist of non-negative terms. Therefore, $\mathrm{CCI}(Z \to \hat{Y}) = 0$, *if and only if* all the terms in the summation are zero, *i.e.*, for all z_1, z_2 and u_x with $\Pr(Z = z_1), \Pr(Z = z_2), \Pr(U_X = u_x) > 0$, $|h(z_1,u_x) - h(z_2,u_x)| = 0$. This is equivalent to $h(z,u_x)$ being constant over all possible values of z with $\Pr(Z = z) > 0$ given a fixed value of u_x, and this should happen over all values of u_x with $\Pr(U_X = u_x)$.

Now, observe that,

$$\begin{aligned} \mathrm{I}(Z; (\hat{Y}, U_X)) &= \mathrm{I}(Z; \hat{Y} \mid U_X) + \mathrm{I}(Z; U_X) && (14)\\ &= \mathrm{I}(Z; \hat{Y} \mid U_X) && [Z \perp\!\!\!\perp U_X] \quad (15)\\ &= \mathrm{H}(\hat{Y} \mid U_X) - \mathrm{H}(\hat{Y} \mid U_X, Z) && \text{[By Definition]} \quad (16)\\ &= \mathrm{H}(\hat{Y} \mid U_X). && [\hat{Y} \text{ is completely determined by } Z \text{ and } U_X] \quad (17) \end{aligned}$$

$\mathrm{H}(\hat{Y} \mid U_X)$ can be 0 *if and only if* $h(z,u_x)$ is constant over all possible values of z with $\Pr(Z = z) > 0$ given a fixed value of u_x, and this should happen over all u_x with $\Pr(U_X = u_x) > 0$. Thus, $\mathrm{CCI}(Z \to \hat{Y}) = 0$ if and only if $\mathrm{I}(Z; (\hat{Y}, U_X)) = 0$. ☐

B. Connections to Counterfactual Fairness

We note that the concept of counterfactual causal influence (often referred to as only "influence") is derived from a separate body of work [63]–[67]) outside the fairness literature. The original definition of counterfactual fairness in [18] was stated differently (without using CCI), although the connection with CCI has been hinted at in [20]. Here, for the sake of completeness, we will formally show in Lemma 6 that $\mathrm{CCI}(Z \to \hat{Y}) = 0$ is equivalent to the counterfactual fairness criterion proposed in [18]. What this means is that, our proposed quantification of total bias is also 0 if and only if a model is counterfactually fair.

First, we clarify the differences in notation between our work and [18]. In our work, $X = f(Z, U_X)$ and $\hat{Y} = r(X) = r \circ f(Z, U_X) = h(Z, U_X)$ where $h = r \circ f$. In [18], $\hat{Y}_{Z \leftarrow z_1}(U)$ denotes the random variable $\hat{Y}$ when the value of Z is fixed as z_1 by an intervention, i.e., $\hat{Y}_{Z \leftarrow z_1}(U) = h(z_1, U_X)$. Alongside, we also clarify that the event that X takes the value x when Z is fixed as z_1 refers to the event that U_X takes a value from the set $\mathcal{S}(x, z_1) = \{u_x : \ x = f(z_1, u_x),\ \Pr(U_X = u_x) > 0\}$ because $X = f(Z, U_X)$.

Definition 8 (Counterfactual Fairness given $X = x$ and $Z = z_1$ [18])**.** *A predictor* $\hat{Y}$ *is counterfactually fair given the protected attribute* $Z = z_1$ *and the observed variable* $X = x$*, if we have,*

$$\begin{aligned} &\Pr(\hat{Y}_{Z \leftarrow z_1}(U) = y \mid X \textit{ takes the value } x \textit{ when } Z \textit{ is fixed as } z_1) \\ &= \Pr(\hat{Y}_{Z \leftarrow z_2}(U) = y \mid X \textit{ takes the value } x \textit{ when } Z \textit{ is fixed as } z_1), \end{aligned} \tag{18}$$

for all attainable y *and* z_2*. In our notations, this definition is equivalent to the following: Given the sensitive attribute* $Z = z_1$ *and the observed variable* $X = x$*,*

$$\Pr(h(z_1, U_X) = y \mid \ U_X \in \mathcal{S}(x, z_1)) = \Pr(h(z_2, U_X)) = y \mid \ \ U_X \in \mathcal{S}(x, z_1)), \tag{19}$$

for all attainable y *and* z_2*, where* $\mathcal{S}(x, z_1) = \{u_x : \ x = f(z_1, u_x),\ \Pr(U_X = u_x) > 0\}$*.*

Next, we show that $\mathrm{CCI}(Z \to \hat{Y}) = 0$ is equivalent to the counterfactual fairness criterion of [18].

Lemma 6. $\mathrm{CCI}(Z \to \hat{Y}) = 0$ *is equivalent to counterfactual fairness (Definition 8) for all* $X = x$ *and* $Z = z_1$ *with* $\Pr(X = x, Z = z_1) > 0$*.*

Proof of Lemma 6. Suppose that, $\mathrm{CCI}(Z \to \hat{Y}) = 0$. Recall from Lemma 1, that $\mathrm{CCI}(Z \to \hat{Y}) = 0$ is equivalent to the criterion that $h(z_1, u_x) = h(z_2, u_x)$ for all attainable z_1, z_2 given a particular value of u_x, and this should hold for all u_x with $\Pr(U_X = u_x) > 0$. Therefore, for any particular $X = x$ and $Z = z_1$ with $\Pr(X = x, Z = z_1) > 0$,

$$\Pr(h(z_1, U_X) = y \mid \; U_X \in \mathcal{S}(x, z_1)) = \Pr(h(z_2, U_X)) = y \mid \quad U_X \in \mathcal{S}(x, z_1)), \tag{20}$$

because $h(z_1, u_x) = h(z_2, u_x)$ for all $u_x \in \mathcal{S}(x, z_1)$. Thus, we show that $\mathrm{CCI}(Z \to \hat{Y}) = 0$ implies counterfactual fairness.

Now, we prove the implication in the other direction. Suppose that the counterfactual fairness criterion (19) holds for all $X = x$ and $Z = z_1$ with $\Pr(X = x, Z = z_1) > 0$.

First consider any particular $X = x$ and $Z = z_1$ with $\Pr(X = x, Z = z_1) > 0$. Since $\Pr(X = x, Z = z_1) > 0$, there exists at least one u_x with $\Pr(U_X = u_x) > 0$ such that $x = f(z_1, u_x)$. So, the set $\mathcal{S}(x, z_1)$ is non-empty. Equation (19) implies that,

$$\Pr(h(z_1, U_X) = y \mid \; U_X \in \mathcal{S}(x, z_1)) = \Pr(h(z_2, U_X)) = y \mid \quad U_X \in \mathcal{S}(x, z_1)) \; \forall \text{ attainable } y, z_2. \tag{21}$$

This leads to,

$$\Pr(h(z_1, U_X) = y, \; U_X \in \mathcal{S}(x, z_1)) = \Pr(h(z_2, U_X) = y, \quad U_X \in \mathcal{S}(x, z_1)) \; \forall \text{ attainable } y, z_2. \tag{22}$$

Or,

$$\sum_{u_x \in \mathcal{S}(x, z_1)} \Pr(U_X = u_x) \mathbb{1}(h(z_1, u_x) = y) = \sum_{u_x \in \mathcal{S}(x, z_1)} \Pr(U_X = u_x) \mathbb{1}(h(z_2, u_x) = y) \; \forall \text{ attainable } y, z_2. \tag{23}$$

Now, observe that, $f(z_1, u_x) = x$ for all $u_x \in \mathcal{S}(x, z_1)$, and thus $h(z_1, u_x) = r \circ f(z_1, u_x)$ takes the same value for all $u_x \in \mathcal{S}(x, z_1)$. Let $h(z_1, u_x) = \tilde{y}$ for all $u_x \in \mathcal{S}(x, z_1)$. Then, for (23) to hold, we need,

$$\sum_{u_x \in \mathcal{S}(x, z_1)} \Pr(U_X = u_x)(1 - \mathbb{1}(h(z_2, u_x) = \tilde{y})) = 0 \; \forall \text{ attainable } z_2.$$

This holds if and only if $\mathbb{1}(h(z_2, u_x) = \tilde{y}) = 1$ for all $u_x \in \mathcal{S}(x, z_1)$ and for all attainable z_2. Thus, the counterfactual fairness criterion (19) *for a particular* $X = x, Z = z_1$ with $\Pr(X = x, Z = z_1) > 0$ implies that for all $u_x \in \mathcal{S}(x, z_1)$,

$$h(z_2, u_x) = h(z_1, u_x) \;\; \forall \text{ attainable } z_2. \tag{24}$$

Because the counterfactual criterion (19) holds for all $X = x, Z = z_1$ with $\Pr(X = x, Z = z_1) > 0$, we therefore have (24) hold for all

$$u_x \in \cup_{\{x, z_1 : \Pr(X = x, Z = z_1) > 0\}} \mathcal{S}(x, z_1).$$

Now, because U_X is independent of Z, for any u_x^* with $\Pr(U_X = u_x^*) > 0$, there always exists some x^* such that $x^* = f(z_1, u_x^*)$, and $\Pr(X = x^*, Z = z_1) \geq \Pr(U_X = u_x^*, Z = z_1) > 0$. Thus, $u_x^* \in S(x^*, z_1)$ for some (x^*, z_1) with $\Pr(X = x^*, Z = z_1) > 0$. Thus,

$$\{u_x : \Pr(U_X = u_x) > 0\} \subseteq \cup_{\{x, z_1 : \Pr(X = x, Z = z_1) > 0\}} \mathcal{S}(x, z_1),$$

implying that $h(z_2, u_x) = h(z_1, u_x)$ for all attainable z_1, z_2 given a particular value of u_x, and this holds for all u_x with $\Pr(U_X = u_x) > 0$. This is equivalent to $\mathrm{CCI}(Z \to \hat{Y}) = 0$ (recall Lemma 1).

□

Appendix B
Relevant Information-Theoretic Properties

Lemma 7 (Conditional DPI)**.** *For all (A, A', B, X_c) such that $(B, X_c) - A - A'$ form a Markov chain, we have the following conditional form of the Data Processing Inequality (DPI):* $\mathrm{I}(A; B \mid X_c) \geq \mathrm{I}(A'; B \mid X_c)$.

Proof of Lemma 7. From the Markov chain, we have $\mathrm{I}(A'; (B, X_c) \mid A) = 0$. Because, $\mathrm{I}(A'; (B, X_c) \mid A) = \mathrm{I}(A'; X_c \mid A) + \mathrm{I}(A'; B \mid A, X_c)$ by chain rule and mutual information is non-negative, we also have $\mathrm{I}(A'; B \mid A, X_c) = 0$. Now, similar to the proof of DPI, we have:

$$\mathrm{I}(A'; B \mid X_c) + \mathrm{I}(A; B \mid A', X_c) = \mathrm{I}(A; B \mid X_c) + \mathrm{I}(A'; B \mid A, X_c) = \mathrm{I}(A; B \mid X_c), \tag{25}$$

because $\mathrm{I}(A'; B \mid A, X_c) = 0$. This leads to $\mathrm{I}(A; B \mid X_c) \geq \mathrm{I}(A'; B \mid X_c)$. □

Lemma 8 (Triangle Inequality of Unique Information)**.** *For all (Z, B, A, X_c), we have:*

$$\mathrm{Uni}(Z : A | X_c) \leq \mathrm{Uni}(Z : A | B) + \mathrm{Uni}(Z : B | X_c).$$

This result is derived in [85, Proposition 2].

Lemma 9 (Monotonicity under local operations on Z)**.** *Let $Z' = f(Z)$ where $f(\cdot)$ is a deterministic function. Then, we have:*

$$\mathrm{Uni}(Z : B|X_c) \geq \mathrm{Uni}(Z' : B|X_c).$$

This result is derived in [73, Lemma 31]. We include a proof for completeness.

Proof of Lemma 9. Let P' be the true joint distribution of (Z', B, X_c) and P be the true joint distribution of (Z, B, X_c). Also let $Q^* = \arg\min_{Q \in \Delta_P} \mathrm{I}_Q(Z; B \mid X_c)$ where Δ_P is the set of all joint distributions of (Z, B, X_c) with the same marginals between (Z, B) and (Z, X_c) as the true joint distribution P. Let us also define

$$Q'^*(z', b, x_c) = \sum_z \Pr(z' \mid z) Q^*(z, b, x_c),$$

where $\Pr(z' \mid z)$ is the true conditional distribution of $Z' = f(Z)$ given Z.

Now, observe that,

$$\begin{aligned}
\mathrm{Uni}(Z : B|X_c) &= \min_{Q \in \Delta_P} \mathrm{I}_Q(Z; B \mid X_c) && \text{[By Definition of Unique Information]} \\
&= \mathrm{I}_{Q^*}(Z; B \mid X_c) && \text{[By Definition of } Q^*\text{]} \\
&\overset{(a)}{\geq} \mathrm{I}_{Q'^*}(Z'; B \mid X_c) \\
&\overset{(b)}{\geq} \min_{Q' \in \Delta_{P'}} \mathrm{I}_{Q'}(Z'; B \mid X_c) \\
&= \mathrm{Uni}(Z' : B|X_c) && \text{[By Definition of Unique Information].}
\end{aligned} \tag{26}$$

Here (a) holds using the conditional form of the Data Processing inequality (Lemma 7) as follows. Consider the random variables (Z, B, X_c) following distribution Q^* and $Z' = f(Z)$. Then, $(B, X_c) - Z - Z'$ form a Markov chain. Also note that (b) holds because Q'^* belongs to $\Delta_{P'}$ which is the set of all joint distributions of (Z', B, X_c) with the same marginals between (Z', B) and (Z', X_c) as the true joint distribution P'. □

Lemma 10 (Monotonicity under local operations on B)**.** *Let $B' = f(B)$ where $f(\cdot)$ is a deterministic function. Then, we have:*

$$\mathrm{Uni}(Z : B|X_c) \geq \mathrm{Uni}(Z : B'|X_c).$$

This result is derived in [73, Lemma 31]. We include a proof for completeness.

Proof of Lemma 10. Let P' be the true joint distribution of (Z, B', X_c) and P be the true joint distribution of (Z, B, X_c). Also let $Q^* = \arg\min_{Q \in \Delta_P} \mathrm{I}_Q(Z; B \mid X_c)$ where Δ_P is the set of all joint distributions of (Z, B, X_c) with the same marginals between (Z, B) and (Z, X_c) as the true joint distribution P. Let us also define

$$Q'^*(z, b', x_c) = \sum_b \Pr(b' \mid b) Q^*(z, b, x_c),$$

where $\Pr(b' \mid b)$ is the true conditional distribution of $B' = f(B)$ given B.

Now, observe that,

$$\begin{aligned}
\mathrm{Uni}(Z : B|X_c) &= \min_{Q \in \Delta_P} \mathrm{I}_Q(Z; B \mid X_c) && \text{[By Definition of Unique Information]} \\
&= \mathrm{I}_{Q^*}(Z; B \mid X_c) && \text{[By Definition of } Q^*\text{]} \\
&\overset{(a)}{\geq} \mathrm{I}_{Q'^*}(Z; B' \mid X_c) \\
&\overset{(b)}{\geq} \min_{Q' \in \Delta_{P'}} \mathrm{I}_{Q'}(Z; B' \mid X_c) \\
&= \mathrm{Uni}(Z : B'|X_c) && \text{[By Definition of Unique Information].}
\end{aligned} \tag{27}$$

Here (a) holds using the conditional form of the Data Processing inequality (Lemma 7) as follows. Consider the random variables (Z, B, X_c) following distribution Q^* and $B' = f(B)$. Then, $(Z, X_c) - B - B'$ form a Markov chain. Also note that (b) holds because Q'^* belongs to $\Delta_{P'}$ which is the set of all joint distributions of (Z, B', X_c) with the same marginals between (Z, B') and (Z, X_c) as the true joint distribution P'. □

Lemma 11 (Monotonicity under adversarial side information)**.** *For all* (A, B, X_c, X'_c)*, we have:*

$$\mathrm{Uni}(A : B|(X_c, X'_c)) \leq \mathrm{Uni}(A : B|X_c).$$

This result is derived in [73, Lemma 32].

Lemma 12 (Maximal conditional mutual information)**.** *Let* $A = f(Z, U_X)$ *where* $Z \perp\!\!\!\perp U_X$ *and* $B = g(U_X)$ *for some deterministic functions* $f(\cdot)$ *and* $g(\cdot)$ *respectively. Then,*

$$\mathrm{I}(Z; A \mid U_X) \geq \mathrm{I}(Z; A \mid B). \tag{28}$$

Proof of Lemma 12. Observe that,

$$\begin{aligned}
&\mathrm{I}(Z; U_X \mid A, B)) \geq 0 && \text{[non-negativity property]}\\
&\implies \mathrm{H}(Z \mid A, B) - \mathrm{H}(Z \mid A, B, U_X) \geq 0 && \text{[by definition]}\\
&\implies \mathrm{H}(Z \mid A, B) - \mathrm{H}(Z \mid A, U_X) \geq 0 && [B = g(U_X)]\\
&\implies \mathrm{H}(Z) - \mathrm{H}(Z \mid A, U_X) \geq \mathrm{H}(Z) - \mathrm{H}(Z \mid A, B) &&\\
&\implies \mathrm{H}(Z \mid U_X) - \mathrm{H}(Z \mid A, U_X) \geq \mathrm{H}(Z \mid B) - \mathrm{H}(Z \mid A, B) && [Z \perp\!\!\!\perp U_X \text{ and } Z \perp\!\!\!\perp B]\\
&\implies \mathrm{I}(Z; A \mid U_X) \geq \mathrm{I}(Z; A \mid B). && (29)
\end{aligned}$$

□

Lemma 13 (Absence of counterfactual causal influence)**.** *Let* $\hat{Y} = h(Z, U_X)$ *where* $Z \perp\!\!\!\perp U_X$ *and* $X_c = g(Z, U_X)$ *for some deterministic functions* $h(\cdot)$ *and* $g(\cdot)$ *respectively. Then* $\mathrm{CCI}(Z \to \hat{Y}) = 0$ *implies* $\mathrm{Uni}(Z : (\hat{Y}, U_X)|X_c) = 0$ *and also* $\mathrm{Uni}(Z : \hat{Y}|X_c) = 0$*.*

Proof of Lemma 13. $\mathrm{CCI}(Z \to \hat{Y}) = 0$ is equivalent to $\mathrm{I}(Z; (\hat{Y}, U_X)) = 0$ (using Lemma 1). Now,

$$\mathrm{Uni}(Z : (\hat{Y}, U_X)|X_c) \overset{(a)}{\leq} \mathrm{I}(Z; (\hat{Y}, U_X)) = 0,$$

where (a) holds from (2) in Section II-A and non-negativity of PID. Also,

$$\mathrm{Uni}(Z : \hat{Y}|X_c) \overset{(a)}{\leq} \mathrm{I}(Z; \hat{Y}) \overset{(b)}{\leq} \mathrm{I}(Z; (\hat{Y}, U_X)) = 0,$$

where (a) holds from (2) in Section II-A and non-negativity of PID terms, and (b) holds from the chain rule and non-negativity of mutual information.

□

Lemma 14 (Zero-synergy property of deterministic functions)**.** *Let* $f(Z)$ *be any deterministic function of* Z*, and let* X_c *be any random variable. Then,*

$$\mathrm{Syn}(Z : (f(Z), X_c)) = \mathrm{Syn}(Z : (X_c, f(Z))) = 0. \tag{30}$$

This leads to $\mathrm{Uni}(Z : f(Z)|X_c) = \mathrm{I}(Z; f(Z)|X_c)$ *and* $\mathrm{Uni}(Z : X_c|f(Z)) = \mathrm{I}(Z; X_c|f(Z))$*.*

Proof of Lemma 14:. Recall from the definition of $\mathrm{Uni}(Z : B|X_c)$ that Δ denotes the set of all joint distributions of (Z, B, X_c) and Δ_p is the set of all such joint distributions that have the same marginals for (Z, B) and (Z, X_c) as the true distribution, *i.e.*,

$$\Delta_p = \{Q \in \Delta : \ q(z, b) = \Pr(Z = z, B = b) \text{ and } q(z, x_c) = \Pr(Z = z, X_c = x_c)\}. \tag{31}$$

We first show that if $B = f(Z)$, then Δ_p is only a singleton set which only consists of the true distribution.

Observe that, for any $Q \in \Delta_p$,

$$\begin{aligned}
q(z,b,x_c) &= q(z)q(b|z)q(x_c|b,z) && \text{[chain rule of probability]}\\
&= \Pr(Z=z)\Pr(B=b|Z=z)q(x_c|b,z) && [q(z,b)=\Pr(Z=z,B=b)]\\
&= \begin{cases}\Pr(Z=z)q(x_c|b,z), & \text{if } b=f(z)\\ 0, & \text{otherwise}\end{cases} && [\Pr(B=b|Z=z)=1 \text{ only if } b=f(z)]\\
&= \begin{cases}\Pr(Z=z)q(x_c|z), & \text{if } b=f(z)\\ 0, & \text{otherwise}\end{cases} && [b \text{ is entirely determined by } z]\\
&= \begin{cases}\Pr(Z=z)\Pr(X_c=x_c|Z=z), & \text{if } y=f(z)\\ 0, & \text{otherwise}\end{cases} && [q(x_c|z)=\Pr(X_c=x_c|Z=z)]\\
&= \Pr(Z=z,B=b,X_c=x_c). && (32)
\end{aligned}$$

Thus, for $B=f(Z)$,

$$\mathrm{Uni}(Z:B|X_c) = \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;B|X_c) = \mathrm{I}(Z;B|X_c). \tag{33}$$

This leads to $\mathrm{Syn}(Z:(f(Z),X_c)) = \mathrm{I}(Z;f(Z)|X_c) - \mathrm{Uni}(Z:f(Z)|X_c) = 0$ (using (3) in Section II-A). Note that, $\mathrm{Syn}(Z:(f(Z),X_c))$ is symmetric between $f(Z)$ and X_c. □

Appendix C
Appendix to Section III

Here, we provide the proofs of the results as well as additional discussion to supplement Section III. For convenience, we repeat the statements of the results.

A. Proof of Theorem 1 and Lemma 2

Theorem 1 (Properties)**.** *Properties 1-6 are satisfied by our proposed measure*

$$M^*_{NE} = \min_{U_a,U_b} \mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c) \textit{ such that } U_a = U_X \backslash U_b.$$

Proof of Theorem 1. Here, we formally show that our proposed measure satisfies all the four desirable properties. We restate each of the properties again and then show that they are is satisfied.

Property 1 (Absence of Exemptions)**.** *If the set of critical features, $X_c=\phi$, then a measure M_{NE} should be equal to the total bias, i.e.,* $\mathrm{I}(Z;(\hat{Y},U_X))$.

When $X_c=\phi$, we have $\mathrm{Uni}(Z,U_a:\hat{Y},U_b|X_c) = \mathrm{I}(Z,U_a;\hat{Y},U_b)$. We are required to show that $\min_{U_a,U_b \text{ s.t. } U_a=U_X\backslash U_b} \mathrm{I}(Z,U_a;\hat{Y},U_b)$ is equal to $\mathrm{I}(Z;(\hat{Y},U_X))$. Note that,

$$\begin{aligned}
\mathrm{I}(Z,U_a;\hat{Y},U_b) &= \mathrm{H}(\hat{Y},U_b) - \mathrm{H}(\hat{Y},U_b \mid Z,U_a) && \text{[By Definition]}\\
&= \mathrm{H}(\hat{Y}\mid U_b) + \mathrm{H}(U_b) - \mathrm{H}(U_b\mid Z,U_a) - \mathrm{H}(\hat{Y}\mid U_b,Z,U_a) && \text{[Chain Rule]}\\
&= \mathrm{H}(\hat{Y}\mid U_b) + \mathrm{H}(U_b) - \mathrm{H}(U_b\mid Z,U_a) && [\hat{Y} \text{ is entirely determined by } Z,U_a,U_b]\\
&= \mathrm{H}(\hat{Y}\mid U_b) && [Z,U_a,U_b \text{ are mutually independent}]\\
&\geq \mathrm{H}(\hat{Y}\mid U_X) && \text{[conditioning reduces entropy]}\\
&= \mathrm{H}(\hat{Y}\mid U_X) - \mathrm{H}(\hat{Y}\mid Z,U_X) + \mathrm{I}(Z;U_X) && [\hat{Y} \text{ entirely determined by } Z,U_X, \text{ and } Z \perp\!\!\!\perp U_X]\\
&= \mathrm{I}(Z;\hat{Y}\mid U_X) + \mathrm{I}(Z;U_X) && \text{[By Definition]}\\
&= \mathrm{I}(Z;(\hat{Y},U_X)). && \text{[By Chain Rule]} \quad (34)
\end{aligned}$$

Thus, $\mathrm{I}(Z,U_a;\hat{Y},U_b) \geq \mathrm{I}(Z;(\hat{Y},U_X))$ with equality when $U_b=U_X, U_a=\phi$.

Property 2 (Non-Increasing with More Exemptions)**.** *For a fixed set of features X and a fixed model $\hat{Y}=h(Z,U_X)$, a measure M_{NE} should be non-increasing if a feature is removed from X_g and added to X_c.*

Let X'_c denote the additional feature that is to be removed from X_g and is to be added to X_c. From Lemma 11, we have,

$$\mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|(X_c,X'_c)) \leq \mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c), \tag{35}$$

for any U_a,U_b. Thus,

$$\min_{U_a,U_b \text{ s.t. } U_a=U_X\backslash U_b} \mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|(X_c,X'_c)) \leq \min_{U_a,U_b \text{ s.t. } U_a=U_X\backslash U_b} \mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c). \tag{36}$$

Property 3 (Complete Exemption). *M_{NE} should be 0 if all features are exempt, i.e., $X_c = X$ and $X_g = \phi$.*

Observe that, when $X = X_c$,

$$\begin{aligned}
M^*_{NE} &= \min_{U_a, U_b \text{ s.t. } U_a = U_X \setminus U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X) \\
&\leq \mathrm{Uni}(Z, U_X : \hat{Y} | X) \\
&\leq \mathrm{I}(Z, U_X ; \hat{Y} \mid X) && \text{[(3) in Section II-A and non-negativity of PID terms]} \\
&= \mathrm{H}(\hat{Y} \mid X) - \mathrm{H}(\hat{Y} \mid Z, U_X, X) && \text{[By Definition]} \\
&= 0. && \text{[$\hat{Y}$ is a deterministic function of X]} \qquad (37)
\end{aligned}$$

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

$$\begin{aligned}
M^*_{NE} &= \min_{U_a, U_b \text{ s.t. } U_a = U_X \setminus U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) \\
&\leq \mathrm{Uni}(Z : (\hat{Y}, U_X) | X_c) \\
&\leq \mathrm{I}(Z; (\hat{Y}, U_X)). \quad \text{[(2) in Section II-A and non-negativity of PID terms]} \qquad (38)
\end{aligned}$$

Thus, $\mathrm{I}(Z; (\hat{Y}, U_X)) = 0$ implies $M_{NE} = 0$.

Property 5 (Non-Exempt Statistically Visible Bias). *M_{NE} should be strictly greater than 0 if $\hat{Y}$ has any unique information about Z. Thus, $\mathrm{Uni}(Z : \hat{Y} | X_c) > 0$ should imply that $M_{NE} > 0$.*

$$\begin{aligned}
M^*_{NE} &= \min_{U_a, U_b \text{ s.t. } U_a = U_X \setminus U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) \\
&= \mathrm{Uni}((Z, U^*_a) : (\hat{Y}, U^*_b) | X_c) && \text{[for some (U^*_a, U^*_b)]} \\
&\geq \mathrm{Uni}(Z : (\hat{Y}, U^*_b) | X_c) && \text{[Using Lemma 9]} \\
&\geq \mathrm{Uni}(Z : \hat{Y} | X_c). && \text{[Using Lemma 10]} \qquad (39)
\end{aligned}$$

Thus, $\mathrm{Uni}(Z : \hat{Y} | X_c) > 0$ implies that $M_{NE} > 0$.

Property 6 (Non-Exempt Masked Bias). *M_{NE} should be non-zero in Example 3, the canonical example of non-exempt masked bias. However, M_{NE} should be 0 if $(Z, U_a) - X_c - (\hat{Y}, U_b)$ form a Markov chain for some subsets $U_a, U_b \subseteq U_X$ such that $U_a = U_X \setminus U_b$.*

First we will show that $M^*_{NE} > 0$ for the canonical example of non-exempt bias where $\hat{Y} = Z \oplus U_{X_1}$ where Z lies in the non-critical/general features and U_{X_1} can be either critical or non-critical.

Case 1: $X_c = U_{X_1}$, $X_g = Z$ and $\hat{Y} = Z \oplus U_{X_1}$ with $Z, U_{X_1} \sim i.i.d.$ Bern(1/2).

We will check the value of $\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c)$ for different choices of U_a to find the minimum.

For $U_a = \phi$ and $U_b = U_{X_1}$, we have

$$\begin{aligned}
\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) &= \mathrm{Uni}(Z : (\hat{Y}, U_{X_1}) | X_c) && \text{[Substituting the variables]} \\
&= \mathrm{I}(Z; (\hat{Y}, U_{X_1})) - \mathrm{Red}(Z : ((\hat{Y}, U_{X_1}), X_c)) && \text{[Using (2) in Section II-A]} \\
&\overset{(a)}{=} \mathrm{I}(Z; (\hat{Y}, U_{X_1})) \\
&= 1 \text{ bit.} && (40)
\end{aligned}$$

Here (a) holds because $\mathrm{Red}(Z : ((\hat{Y}, U_{X_1}), X_c)) \leq \mathrm{I}(Z; X_c)$ (using (2) in Section II-A and non-negativity of PID terms), and here $\mathrm{I}(Z; X_c) = 0$.

For $U_a = U_{X_1}$ and $U_b = \phi$, we have

$$\begin{aligned}
\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) &= \mathrm{Uni}((Z, U_{X_1}) : \hat{Y} | X_c) && \text{[Substituting the variables]} \\
&= \mathrm{I}((Z, U_{X_1}); \hat{Y} \mid X_c) && \text{[Lemma 14 as $\hat{Y}$ is deterministic function of $f(Z, U_{X_1})$]} \\
&= 1 \text{ bit.} && (41)
\end{aligned}$$

Thus,

$$M^*_{NE} = \min_{U_a, U_b \text{ s.t. } U_a = U_X \setminus U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b) | X_c) = 1 \text{ bit},$$

which is strictly greater than 0.

Case 2: $X_c = \phi$, $X_g = (Z, U_{X_1})$ and $\hat{Y} = Z \oplus U_{X_1}$ with $Z, U_{X_1} \sim i.i.d.$ Bern(1/2).
Since $X_c = \phi$, we can use Property 1 (proved above) to compute

$$M^*_{NE} = \mathrm{I}(Z; (\hat{Y}, U_X)) = 1 \text{ bit},$$

which is strictly greater than 0. Thus, our proposed measure is non-zero in the canonical example of non-exempt masked bias. Now, we move on to the proof of the next part of this property.

Suppose that $(Z, U_a) - X_c - (\hat{Y}, U_b)$ form a Markov chain for some subsets $U_a, U_b \subseteq U_X$ such that $U_a = U_X \backslash U_b$. Then, $\mathrm{I}((Z, U_a); (\hat{Y}, U_b) \mid X_c) = 0$, implying that $\mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b)|X_c) = 0$ for those subsets $U_a, U_b \subseteq U_X$ because unique information is a sub-component of conditional mutual information. Therefore,

$$M^*_{NE} = \min_{U_a, U_b \text{ s.t. } U_a = U_X \backslash U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b)|X_c) \leq 0.$$

Again, using the fact that unique information is non-negative, we have,

$$M^*_{NE} = \min_{U_a, U_b \text{ s.t. } U_a = U_X \backslash U_b} \mathrm{Uni}((Z, U_a) : (\hat{Y}, U_b)|X_c) \geq 0.$$

Thus, $M^*_{NE} = 0$.

□

Lemma 2. *The Markov chain $(Z, U_a) - X_c - (\hat{Y}, U_b)$ implies that the following Markov chains also hold: (i) $Z - X_c - \hat{Y}$; (ii) $(Z, U_a) - X_c - \hat{Y}$; and (ii) $Z - X_c - (\hat{Y}, U_b)$.*

Proof of Lemma 2. We note that the terms $\mathrm{I}(Z; \hat{Y} \mid X_c)$, $\mathrm{I}(Z; (\hat{Y}, U_b) \mid X_c)$ and $\mathrm{I}((Z, U_a); \hat{Y} \mid X_c)$ are all less than or equal to $\mathrm{I}((Z, U_a); (\hat{Y}, U_b) \mid X_c)$ using the chain rule and non-negativity of conditional mutual information.
Thus, if $\mathrm{I}((Z, U_a); (\hat{Y}, U_b) \mid X_c) = 0$, then all those three terms are also 0. □

B. Supporting Derivations

Here, we include the supporting derivations for some of our statements in Section III-A and Section III-B.

Supporting Derivation 1: $\mathrm{Uni}(Z : \hat{Y}|X_c) > 0$ **for Example 5 (discrimination in admissions).**

Proof. Recall that for this example, $X_c = U_{X_1}$, $X_g = Z \oplus U_{X_2}$, and $\hat{Y} = U_{X_1} + Z + U_{X_2}$ with $Z, U_{X_1}, U_{X_2} \sim$ i.i.d. Bern(1/2). The claim can be verified as follows:

$$\begin{aligned}
\mathrm{Uni}(Z : \hat{Y}|X_c) &= \mathrm{I}(Z; \hat{Y}) - \mathrm{Red}(Z : (\hat{Y}, X_c)) && \text{[using (2) in Section II-A]} \\
&\overset{(a)}{\geq} \mathrm{I}(Z; \hat{Y}) - \mathrm{I}(Z; X_c) \\
&\overset{(b)}{=} \mathrm{I}(Z; \hat{Y}) \\
&\overset{(c)}{>} 0,
\end{aligned}$$

where (a) holds because $\mathrm{Red}(Z : (\hat{Y}, X_c)) \leq \mathrm{I}(Z; X_c)$ (using (2) in Section II-A and non-negativity of all PID terms) and (b) holds because $\mathrm{I}(Z; X_c) = 0$. Lastly, (c) holds because $\hat{Y}$ and Z are not independent of each other for this specific example. □

Supporting Derivation 2: $\mathrm{Uni}(Z : \hat{Y}|X_c) > 0$ **for Counterexample 3 (discrimination by unmasking).**

Proof. Recall that for this example, $X_c = Z \oplus U_{X_1}$, $X_g = U_{X_1}$ and $\hat{Y} = Z$ with $Z, U_{X_1} \sim$ i.i.d. Bern(1/2).
The claim can be verified as follows:

$$\begin{aligned}
\mathrm{Uni}(Z : \hat{Y}|X_c) &= \mathrm{I}(Z; \hat{Y}) - \mathrm{Red}(Z : (\hat{Y}, X_c)) && \text{[using (2) in Section II-A]} \\
&\overset{(a)}{\geq} \mathrm{I}(Z; \hat{Y}) - \mathrm{I}(Z; X_c) \\
&\overset{(b)}{=} 1 \text{ bit},
\end{aligned}$$

where (a) holds because $\mathrm{Red}(Z : (\hat{Y}, X_c)) \leq \mathrm{I}(Z; X_c)$ (using (2) in Section II-A and non-negativity of all PID terms) and (b) holds because $\mathrm{I}(Z; X_c) = 0$. □

Supporting Derivation 3: $\mathrm{Uni}(Z : (\hat{Y}, U_X)|X_c) > 0$ **in Counterexample 4 and Example 4.**

Proof. First consider Counterexample 4.

$$\begin{aligned}
&\mathrm{Uni}(Z:(\hat{Y},U_X)|X_c) = \mathrm{Uni}(Z:(Z+U_{X_1},U_X)|Z+U_{X_1}) && \text{[Substituting the variables]}\\
&\overset{(a)}{\geq} \mathrm{Uni}(Z:Z|Z+U_{X_1})\\
&\overset{(b)}{=} \mathrm{I}(Z;Z \mid Z+U_{X_1})\\
&\overset{(c)}{>} 0.
\end{aligned}$$

Here, (a) holds because Z is a deterministic function of $(Z+U_{X_1},U_X)$ and unique information is non-increasing under local operations of B (see Lemma 10 in Appendix B). Next, (b) holds because if we consider Δ_p, the set of joint distributions of $(Z,Z,Z+U_{X_1})$, such that the marginals (Z,Z) and $(Z,Z+U_{X_1})$ are the same as the marginals of the true joint distribution, we find that there is only one distribution in this set, which is exactly the true distribution. Thus, $\mathrm{Uni}(Z:Z|Z+U_{X_1}) = \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;Z \mid Z+U_{X_1}) = \mathrm{I}(Z;Z \mid Z+U_{X_1})$. Lastly (c) holds because,

$$\mathrm{I}(Z;Z \mid Z+U_{X_1}) = \mathrm{H}(Z|Z+U_{X_1}) - \mathrm{H}(Z|Z,Z+U_{X_1}) = \mathrm{H}(Z|Z+U_{X_1}) = \sum_{t=0,1,2} \mathrm{H}(Z|Z+U_{X_1}=t)\Pr(Z+U_{X_1}{=}t).$$

Using the fact that $Z,U_{X_1} \sim i.i.d.$ Bern(1/2), we can compute $\mathrm{H}(Z|Z+U_{X_1}=0)=0$, $\mathrm{H}(Z|Z+U_{X_1}=1)=h_b(1/2)=1$, and $\mathrm{H}(Z|Z+U_{X_1}=2)=0$. Here, $h_b(\cdot)$ is the binary entropy function [79] given by $h_b(p) = -p\log_2(p)-(1-p)\log_2(1-p)$. Also note that, $\Pr(Z+U_{X_1}=1)=1/2$. So, $\mathrm{I}(Z;Z \mid Z+U_{X_1}) = 0.5$ bits.

Next, consider Example 4. The derivation is similar as above because Z can be obtained from local operations on $(\hat{Y},U_X)$.

$$\begin{aligned}
&\mathrm{Uni}(Z:(\hat{Y},U_X)|X_c) = \mathrm{Uni}(Z:(Z+U_{X_1}+U_{X_2},U_X)|Z+U_{X_1}) && \text{[Substituting the variables]}\\
&\geq \mathrm{Uni}(Z:Z|Z+U_{X_1})\\
&= \mathrm{I}(Z;Z \mid Z+U_{X_1}) > 0.
\end{aligned}$$

□

Supporting Derivation 4: Exact computation of $\mathrm{Uni}(Z:\hat{Y}|X_c)$ and M^*_{NE} for Example 5.

$$\begin{aligned}
\mathrm{Uni}(Z:\hat{Y}|X_c) &\overset{(a)}{=} \mathrm{I}(Z;\hat{Y})\\
&= \mathrm{H}(Z) - \mathrm{H}(Z|\hat{Y})\\
&= \mathrm{H}(Z) - \mathrm{H}(Z|U_{X_1}+Z+U_{X_2})\\
&= \mathrm{H}(Z) - \sum_{t=0,1,2,3} \mathrm{H}(Z|U_{X_1}+Z+U_{X_2}=t)\Pr(U_{X_1}+Z+U_{X_2}=t)\\
&\overset{(b)}{=} 1 - 3/4\, h_b(1/3) \text{ bits.}
\end{aligned} \tag{42}$$

Here (a) holds because $\mathrm{I}(Z;U_{X_1})=0$, implying $\mathrm{Red}(Z:(\hat{Y},U_{X_1}))=0$ as well (using (2) in Section II-A and non-negativity of PID terms). Lastly, (b) holds because $Z,U_{X_1},U_{X_2} \sim i.i.d.$ Bern(1/2). So, we can exactly compute $\mathrm{H}(Z|U_{X_1}+Z+U_{X_2}=0)=0$, $\mathrm{H}(Z|U_{X_1}+Z+U_{X_2}=1)=h_b(1/3)$, $\mathrm{H}(Z|U_{X_1}+Z+U_{X_2}=2)=h_b(1/3)$, and $\mathrm{H}(Z|U_{X_1}+Z+U_{X_2}=3)=0$. Here, $h_b(\cdot)$ is the binary entropy function [79] given by $h_b(p) = -p\log_2(p)-(1-p)\log_2(1-p)$. Also note that, $\Pr(U_{X_1}+Z+U_{X_2}=1) = \Pr(U_{X_1}+Z+U_{X_2}=2) = 3/8$.

Now, we will examine the value of $\mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c)$ for different choices of U_a to find the minimum.

Let $U_a=\phi$ (and $U_b=U_X$). Then,

$$\begin{aligned}
&\mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c) = \mathrm{Uni}(Z:(\hat{Y},U_{X_1},U_{X_2})|U_{X_1})\\
&\overset{(a)}{=} \mathrm{I}(Z;U_{X_1}+Z+U_{X_2},U_{X_1},U_{X_2})\\
&= \mathrm{I}(Z;U_{X_1},U_{X_2}) + \mathrm{I}(Z;U_{X_1}+Z+U_{X_2} \mid U_{X_1},U_{X_2}) && \text{[Chain Rule]}\\
&= \mathrm{I}(Z;U_{X_1}+Z+U_{X_2} \mid U_{X_1},U_{X_2}) && [Z \text{ is independent of } U_{X_1},U_{X_2}]\\
&= \mathrm{H}(U_{X_1}+Z+U_{X_2} \mid U_{X_1},U_{X_2}) - \mathrm{H}(U_{X_1}+Z+U_{X_2} \mid Z,U_{X_1},U_{X_2}) && \text{[By Definition]}\\
&= \mathrm{H}(U_{X_1}+Z+U_{X_2} \mid U_{X_1},U_{X_2}) && \text{[Deterministic Function]}\\
&= \sum_{u_1,u_2\in\{0,1\}} \mathrm{H}(U_{X_1}+Z+U_{X_2} \mid U_{X_1}=u_1,U_{X_2}=u_2)\Pr(U_{X_1}=u_1,U_{X_2}=u_2)\\
&= \sum_{u_1,u_2\in\{0,1\}} h_b(1/2)\Pr(U_{X_1}=u_1,U_{X_2}=u_2)\\
&= 1 \text{ bit.}
\end{aligned} \tag{43}$$

Here (a) holds again because $\mathrm{I}(Z;U_{X_1})=0$, implying the redundant information is 0 as well (using (2) in Section II-A).

Next, for $U_a=U_{X_2}$ (and $U_b=U_{X_1}$), we have,

$$\begin{aligned}
&\mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c)=\mathrm{Uni}((Z,U_{X_2}):(\hat{Y},U_{X_1})|U_{X_1})\\
&\overset{(a)}{=}\mathrm{I}((Z,U_{X_2});(\hat{Y},U_{X_1}))\\
&=\mathrm{I}((Z,U_{X_2});U_{X_1})+\mathrm{I}((Z,U_{X_2});\hat{Y}\mid U_{X_1}) && \text{[Chain Rule]}\\
&=\mathrm{I}((Z,U_{X_2});\hat{Y}\mid U_{X_1}) && [Z,U_{X_2}\text{ is independent of }U_{X_1}]\\
&=\mathrm{H}(U_{X_1}+Z+U_{X_2}\mid U_{X_1})-\mathrm{H}(U_{X_1}+Z+U_{X_2}\mid U_{X_1},(Z,U_{X_2})) && \text{[By Definition]}\\
&=\mathrm{H}(U_{X_1}+Z+U_{X_2}\mid U_{X_1}) && \text{[Deterministic Function]}\\
&=\sum_{u_1=0,1}\mathrm{H}(U_{X_1}+Z+U_{X_2}\mid U_{X_1}=u_1)\Pr(U_{X_1}=u_1)\\
&={}^1\!/\!_4\log_2 4+{}^1\!/\!_2\log_2 2+{}^1\!/\!_4\log_2 4\\
&={}^3\!/\!_2\text{ bit.}
\end{aligned}\tag{44}$$

Here (a) holds again because $\mathrm{I}((Z,U_{X_2});U_{X_1})=0$, implying the redundant information is 0 as well (using (2) in Section II-A).

Next, for $U_a=U_{X_1}$ (and $U_b=U_{X_2}$), we have,

$$\begin{aligned}
&\mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c)=\mathrm{Uni}((Z,U_{X_1}):(\hat{Y},U_{X_2})|U_{X_1})\\
&\overset{(b)}{=}\mathrm{I}((Z,U_{X_1});(\hat{Y},U_{X_2})\mid U_{X_1})\\
&=\mathrm{I}((Z,U_{X_1});U_{X_2}\mid U_{X_1})+\mathrm{I}((Z,U_{X_1});\hat{Y}\mid U_{X_1},U_{X_2}) && \text{[Chain Rule]}\\
&=\mathrm{I}((Z,U_{X_1});\hat{Y}\mid U_{X_1},U_{X_2}) && \text{[Mutual Independence]}\\
&=\mathrm{H}(\hat{Y}\mid U_{X_1},U_{X_2})-\mathrm{H}(\hat{Y}\mid (Z,U_{X_1}),U_{X_1},U_{X_2}) && \text{[By Definition]}\\
&=\mathrm{H}(\hat{Y}\mid U_{X_1},U_{X_2}) && \text{[Deterministic Function]}\\
&=\mathrm{H}(U_{X_1}+Z+U_{X_2}\mid U_{X_1},U_{X_2})\\
&=1\text{ bit.}
\end{aligned}\tag{45}$$

Here (b) holds because $\mathrm{Syn}((Z,U_{X_1}):(A,B))=0$ if one of the terms A or B is a deterministic function of (Z,U_{X_1}) (using Lemma 14 in Appendix B) and hence unique information becomes equal to the conditional mutual information (see (3) in Section II-A).

Lastly, for $U_a=U_X$ (and $U_b=\phi$), we have,

$$\begin{aligned}
&\mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c)=\mathrm{Uni}((Z,U_{X_1},U_{X_2}):\hat{Y}|U_{X_1})\\
&\overset{(b)}{=}\mathrm{I}((Z,U_{X_1},U_{X_2});\hat{Y}\mid U_{X_1})\\
&=\mathrm{H}(\hat{Y}\mid U_{X_1})-\mathrm{H}(\hat{Y}\mid (Z,U_{X_1},U_{X_2}),U_{X_1}) && \text{[By Definition]}\\
&=\mathrm{H}(\hat{Y}\mid U_{X_1}) && \text{[Deterministic Function]}\\
&={}^1\!/\!_4\log_2 4+{}^1\!/\!_2\log_2 2+{}^1\!/\!_4\log_2 4\\
&={}^3\!/\!_2\text{ bit.}
\end{aligned}\tag{46}$$

Here (b) holds again using Lemma 14 in Appendix B.

Thus, we obtain that,

$$M^*_{NE}=\min_{U_a,U_b\text{ s.t. }U_a=U_X\setminus U_b}\mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c)=1\text{ bit.}$$

This is strictly greater than $\mathrm{Uni}(Z:\hat{Y}|X_c)=1-\frac{3}{4}h_b({}^1\!/\!_3)$ bits, accounting for both non-exempt statistically visible and non-exempt masked biases.

C. Discussion on Other Candidate Measures

Why the product of the two measures $\mathrm{I}(Z;\hat{Y}\mid X_c)$ and $\mathrm{I}(Z;(\hat{Y},U_X))$ does not work?

One might recall that the measure $\mathrm{I}(Z;\hat{Y}\mid X_c)$ resolved most of the examples except in Counterexample 1 where the output $\hat{Y}$ had no counterfactual causal influence of Z and yet this measure gave a false positive conclusion about non-exempt bias. This leads us to examine another candidate measure, i.e., product of $\mathrm{I}(Z;\hat{Y}\mid X_c)$ and $\mathrm{I}(Z;(\hat{Y},U_X))$ where the latter is always 0 whenever there is no counterfactual causal influence of Z on $\hat{Y}$.

Candidate Measure 5. $M_{NE}=\mathrm{I}(Z;\hat{Y}\mid X_c)\times\mathrm{I}(Z;(\hat{Y},U_X))$.

Counterexample 5. *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. However, both $\mathrm{I}(Z;(\hat{Y}, U_X))$ and $\mathrm{I}(Z;\hat{Y} \mid X_c)$ are non-zero for this example. This leads us to examine another candidate measure, which is essentially the common information-theoretic volume between $\mathrm{I}(Z;(\hat{Y}, U_X))$ and $\mathrm{I}(Z;\hat{Y} \mid X_c)$, i.e., a measure of the common reason that can make both $\mathrm{I}(Z;(\hat{Y}, U_X)) > 0$ and $\mathrm{I}(Z;\hat{Y} \mid X_c) > 0$ (overlapping volume).

Measure proposed in [1]: Information-theoretic sub-volume of the intersection between $\mathrm{I}(Z;\hat{Y} \mid X_c)$ and $\mathrm{I}(Z;(\hat{Y}, U_X))$:
The previous counterexample demonstrates that both these measures $\mathrm{I}(Z;\hat{Y} \mid X_c)$ and $\mathrm{I}(Z;(\hat{Y}, U_X))$ can be non-zero for different reasons leading to a false positive conclusion using Candidate Measure 5. Intuitively, we need to identify the common reason that makes them non-zero, if any. This motivates us to examine another candidate (Candidate Measure 6) which is the information-theoretic sub-volume of the intersection between these two measures, as shown in Fig. 13.

Candidate Measure 6. $M_{NE} = \mathrm{Uni}(Z : (\hat{Y}, U_X)|X_c) - \mathrm{Uni}(Z : (\hat{Y}, U_X)|(X_c, \hat{Y}))$.

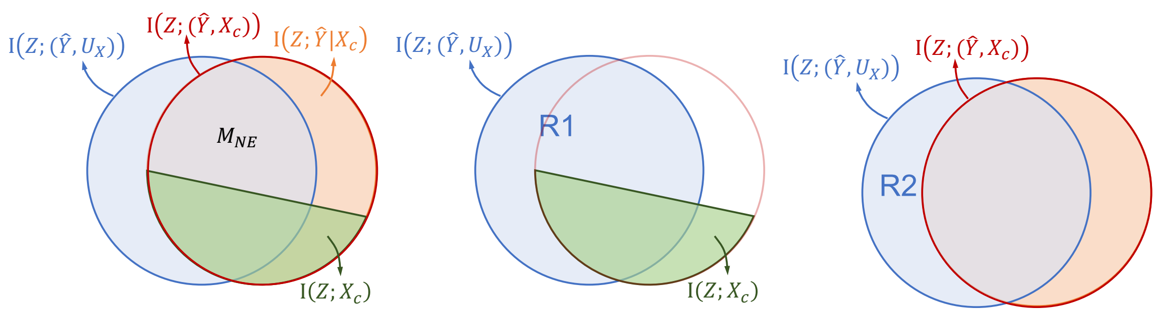


Fig. 13: (Left) Notice that the blue full-circle denotes $\mathrm{I}(Z;(\hat{Y}, U_X))$ and the red full-circle denotes $\mathrm{I}(Z;(\hat{Y}, X_c))$. The term $\mathrm{I}(Z;(\hat{Y}, X_c))$ is equal to the sum of $\mathrm{I}(Z;X_c)$ (green half-circle) and $\mathrm{I}(Z;\hat{Y} \mid X_c)$ (orange half-circle). The candidate measure (M_{NE}) is the intersecting volume between $\mathrm{I}(Z;(\hat{Y}, U_X))$ and $\mathrm{I}(Z;\hat{Y} \mid X_c)$. Next, we show pictorially that this intersecting volume is given by $R1 - R2$ where $R1$ is shown in the middle figure and $R2$ is shown in the rightmost figure. (Middle) Notice that $R1 = \mathrm{Uni}(Z : (\hat{Y}, U_X)|X_c)$. (Right) Notice that $R2 = \mathrm{Uni}(Z : (\hat{Y}, U_X)|(\hat{Y}, X_c))$.

Limitations of Candidate Measure 6: This measure does resolve many of the examples and satisfies several desirable properties (discussed more in [1]). However, it fails to capture certain types of non-exempt masked bias when the mask arises from X_g, e.g., scenarios like Example 6 in Section III-B, where non-exempt masked bias is present even though $Z - X_c - \hat{Y}$ form a Markov chain.

Appendix D
Appendix to Section IV

A. Proof of Theorem 2 and Lemma 3

Theorem 2 (Non-negative Decomposition of Total Bias)**.** *The total bias can be decomposed into four components as follows:*

$$\mathrm{I}(Z;(\hat{Y}, U_X)) = M_{V,NE} + M_{V,E} + M_{M,NE} + M_{M,E}. \tag{11}$$

Here $M_{V,NE} = \mathrm{Uni}(Z : \hat{Y}|X_c)$ and $M_{V,E} = \mathrm{Red}(Z : (\hat{Y}, X_c))$. These two terms add to form $\mathrm{I}(Z;\hat{Y})$ which is the total statistically visible bias. Next, $M_{M,NE} = M^*_{NE} - M_{V,NE}$ where M^*_{NE} is our proposed measure of non-exempt bias (Definition 7), and $M_{M,E} = \mathrm{I}(Z;\hat{Y}, U_X) - \mathrm{I}(Z;\hat{Y}) - M_{M,NE}$. All of these components are non-negative.

Proof of Theorem 2. First consider $M_{V,NE} = \mathrm{Uni}(Z : \hat{Y}|X_c)$ and $M_{V,E} = \mathrm{Red}(Z : (\hat{Y}, X_c))$. Because all PID terms are non-negative by definition, both $M_{V,NE}$ and $M_{V,E}$ are non-negative.

Now, consider $M_{M,E}$. Observe that,

$$\begin{aligned}
M_{M,E} &= \mathrm{I}(Z;(\hat{Y},U_X)) - \mathrm{I}(Z;\hat{Y}) - M_{M,NE} && \text{[By Definition]}\\
&= \mathrm{I}(Z;\hat{Y}) + \mathrm{I}(Z;U_X \mid \hat{Y}) - \mathrm{I}(Z;\hat{Y}) - M_{M,NE} && \text{[Chain Rule for mutual information]}\\
&= \mathrm{I}(Z;U_X \mid \hat{Y}) - M_{M,NE} &&\\
&= \mathrm{I}(Z;U_X \mid \hat{Y}) - M^*_{NE} + M_{V,NE} && \text{[By Definition]}\\
&= \mathrm{I}(Z;U_X \mid \hat{Y}) - \min_{U_a,U_b \text{ s.t. } U_a=U_X\setminus U_b} \mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c) + \mathrm{Uni}(Z:\hat{Y}|X_c) && \text{[By Definition]}\\
&\geq \mathrm{I}(Z;U_X \mid \hat{Y}) - \mathrm{Uni}(Z:(\hat{Y},U_X)|X_c) + \mathrm{Uni}(Z:\hat{Y}|X_c) &&\\
&\geq \mathrm{I}(Z;U_X \mid \hat{Y}) - \mathrm{Uni}(Z:(\hat{Y},U_X)|\hat{Y}) && \text{[Triangle Inequality (Lemma 8)]}\\
&\geq \mathrm{I}(Z;U_X \mid \hat{Y}) - \mathrm{I}(Z;(\hat{Y},U_X) \mid \hat{Y}) && \text{[(3) in Section II-A]}\\
&= \mathrm{I}(Z;U_X \mid \hat{Y}) - \mathrm{I}(Z;U_X \mid \hat{Y}) - \mathrm{I}(Z;\hat{Y} \mid U_X,\hat{Y}) && \text{[Chain Rule for mutual information]}\\
&= 0. && (47)
\end{aligned}$$

Lastly, we consider $M_{M,NE} = \min_{U_a,U_b \text{ s.t. } U_a=U_X\setminus U_b} \mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c) - \mathrm{Uni}(Z:\hat{Y}|X_c)$.

$$\begin{aligned}
M_{NE} &= \min_{U_a,U_b \text{ s.t. } U_a=U_X\setminus U_b} \mathrm{Uni}((Z,U_a):(\hat{Y},U_b)|X_c) - \mathrm{Uni}(Z:\hat{Y}|X_c) &&\\
&= \mathrm{Uni}((Z,U_a^*):(\hat{Y},U_b^*)|X_c) - \mathrm{Uni}(Z:\hat{Y}|X_c) && \text{[for some } (U_a^*,U_b^*)]\\
&\geq \mathrm{Uni}(Z:(\hat{Y},U_b^*)|X_c) - \mathrm{Uni}(Z:\hat{Y}|X_c) && \text{[Using Lemma 9]}\\
&\geq \mathrm{Uni}(Z:\hat{Y}|X_c) - \mathrm{Uni}(Z:\hat{Y}|X_c) && \text{[Using Lemma 10]}\\
&= 0. && (48)
\end{aligned}$$

□

Lemma 3 (Conditioning to Capture Masked Bias)**.** *The following two statements are equivalent:*
- *Masked bias* $\mathrm{I}(Z;(\hat{Y},U_X)) - \mathrm{I}(Z;\hat{Y}) > 0$.
- $\exists$ *a random variable* G *of the form* $G = g(U_X)$ *such that* $\mathrm{I}(Z;\hat{Y} \mid G) - \mathrm{I}(Z;\hat{Y}) > 0$.

Proof of Lemma 3. Before proceeding, note that, $\mathrm{I}(Z;\hat{Y},U_X) = \mathrm{I}(Z;U_X) + \mathrm{I}(Z;\hat{Y} \mid U_X) = \mathrm{I}(Z;\hat{Y} \mid U_X)$ because Z is independent of U_X. This also leads to the masked bias being equal to $\mathrm{I}(Z;\hat{Y} \mid U_X) - \mathrm{I}(Z;\hat{Y})$.

First, we show that the first statement implies the second statement. Suppose that, masked bias $\mathrm{I}(Z;\hat{Y} \mid U_X) - \mathrm{I}(Z;\hat{Y}) > 0$. Then, we can choose the function $G = U_X$ such that $\mathrm{I}(Z;\hat{Y} \mid G) - \mathrm{I}(Z;\hat{Y}) > 0$. Thus, the implication holds.

We will now show that the second statement also implies the first statement. First note that, using Lemma 12, for any deterministic $g(\cdot)$, we always have $\mathrm{I}(Z;\hat{Y} \mid U_X) \geq \mathrm{I}(Z;\hat{Y} \mid g(U_X))$. Now, suppose there exists a $G = g(U_X)$ such that $\mathrm{I}(Z;\hat{Y} \mid G) > \mathrm{I}(Z;\hat{Y})$. Then, $\mathrm{I}(Z;\hat{Y} \mid U_X) \geq \mathrm{I}(Z;\hat{Y} \mid g(U_X)) > \mathrm{I}(Z;\hat{Y})$, implying masked bias is present.

Thus, we prove that the first and second statements are equivalent.

□

Appendix E
Appendix to Section VI

Lemma 4. *[Fairness Properties of* $\mathrm{Uni}(Z:\hat{Y}|X_c)$*] The measure* $\mathrm{Uni}(Z:\hat{Y}|X_c)$ *satisfies Properties 2, 3, 4, and 5.*

Proof of Lemma 4. Property 2 is satisfied using Lemma 11 in Appendix B (originally derived in [73, Lemma 32]).

Property 3 is satisfied because $\hat{Y}$ is a deterministic function of the entire X, and hence the Markov chain $Z - X - \hat{Y}$ holds. Thus $\mathrm{I}(Z;\hat{Y} \mid X_c) = 0$, also implying $\mathrm{Uni}(Z:\hat{Y}|X_c) = 0$.

Property 5 is trivially satisfied because the property itself requires that $\mathrm{Uni}(Z:\hat{Y}|X_c) > 0$.

For Property 4, observe that,

$$\begin{aligned}
&\mathrm{CCI}(Z \to \hat{Y}) = 0 &&\\
&\implies \mathrm{I}(Z;\hat{Y}) = 0 &&\\
&\implies \mathrm{Uni}(Z:\hat{Y}|X_c) + \mathrm{Red}(Z:(\hat{Y},X_c)) = 0 && \text{[Using (2) in Section II-A]}\\
&\implies \mathrm{Uni}(Z:\hat{Y}|X_c) = 0 && \text{[Non-negativity of PID terms].} \quad (49)
\end{aligned}$$

□

Lemma 5. *[Fairness Properties of* $\mathrm{I}(Z;\hat{Y} \mid X_c)$*] The measure* $\mathrm{I}(Z;\hat{Y} \mid X_c)$ *satisfies Properties 3 and 5.*

Proof of Lemma 5. Property 3 is satisfied because $\hat{Y}$ is a deterministic function of the entire X, and hence the Markov chain $Z - X - \hat{Y}$ holds.

For Property 5, observe that

$$\begin{aligned} & \mathrm{Uni}(Z : \hat{Y} | X_c) > 0 \\ \implies & \mathrm{I}(Z; \hat{Y} \mid X_c) > 0 \qquad \text{[Using (3) in Section II-A and non-negativity of PID terms].} \end{aligned} \tag{50}$$

□