ON THE AXIOMS OF INDIRECT INFLUENCE IN MACHINE LEARNING

HAVERFORD INDIRECT INFLUENCE AXIOM GROUP

1. Introduction

A growing stream of work in machine learning focuses on the question of *feature influence*; that is, the goal to measure the impact of a feature on the prediction of a model. In algorithmic fairness, we are interested in whether a model discriminates based on an observed *protected feature* (e.g., race, gender) by describing how a prediction might change if an individual's protected class were to change. Mirroring the multitude of definitions of fairness, there are a wide range of methods for measuring feature influence, which can be broadly categorized into two groups: *direct influence* and *indirect influence.* A *direct influence* method only applies to features fed in the model (as inputs). A feature A has a high direct influence if the value taken by A is heavily relied upon by the model making a prediction. For example, these can be features with large weights (in terms of absolute value) in logistic regression. However, the predictions of a model are not only determined by the features given as inputs, they may also be affected by others that are correlated with some proxy variables in the inputs. For example, zip code is often related to race and may act as a proxy for race in a model. Under this consideration, *indirect influence* method applies to any feature in the space and considers whether a model uses a proxy for A to make its predictions. While a model, which only relies on zip code to make its predictions, exhibits low *direct influence* for race, it is likely that race has a high *indirect influence* since its information can be accessed via the zip code feature.

Direct influence methods can be very useful for model interpretation and enjoy wide applicability. However, indirect influence is often more appropriate in the context of evaluating fairness since sensitive attributes such as race are often deeply interwoven with other societal factors such as socioeconomic status, work history, criminal justice and educational opportunity – ignoring these latent relationships blinds one to discrimination mediated by these proxy variables. Consider, for example, the task of predicting high school GPA. A highly simplified model of variables associated with high school GPA is shown in Figure 1. It is clear that probing only the direct effects of `RACE` on `GPA` along `RACE` → `GPA` path is not sufficient. For example, suppose students from a predominantly black high school had low `GPA`s while students from a predominantly white high school had high GPAs. Then, a model might systematically predict a lower GPA for black students versus white students without a direct access to the RACE feature. Such a model operates along the RACE → STATE OF RESIDENCE → SCHOOL → GPA path, and is likely to include strong dependencies between RACE and predicted GPA. Hence, it is important to consider the influence along all paths starting at RACE and ending at GPA.

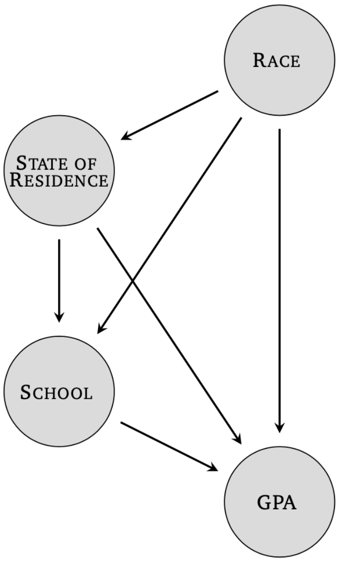


Figure 1. A highly simplified model that represents potential relationships between variables associated with high school GPA.

Recent works on computing indirect influence can be broadly categorized in 4 groups:

- perturbing data to obscure protected information (Feldman et al. 2015; Adebayo and Kagal 2016),
- monitoring the correlation of a protected feature with intermediate representations of a model (Kim et al. 2018),
- modeling feature dependencies to estimate indirect influence directly (Marx et al. 2019) and,
- computing indirect influence based on causality/causal graphs (Farbmacher et al. 2020; Kilbertus et al. 2017).

Each method has inherent strengths and limitations. For comparison, we require a clearly stated desiderata, yet a formal notion of indirect influence has never been proposed in previous works. It is widely acknowledged that success in prediction tasks lies in generalization performance – to see if a learning algorithm is effective, we check the test set performance. However, success in indirect influence measurement is difficult to verify. When two methods give conflicting influence scores, which should we trust? In this paper, we seek to formalize intuitions of indirect influence (e.g. when it should be zero, when it should be the same, etc.). In section 3, we introduce a list of desirable axioms for indirect influence measures. We shall see from section 4 that all these properties can be covered by *efficiency*. To apply these axioms, we detail how prevailing indirect influence methods fail to satisfy some of the properties in section 5. In contrast, we propose in section 6 that *mutual information* (in information theory) can be a measure that meets all requirements. A discussion of our work is given in the end.

2. Related Work

Our work can be broadly considered as an axiomatic approach to *feature influence*. Previous works in this area include Sliwinski et al. (2017), which proposes a set of properties influence measures should satisfy. As a consequence, they characterize the family of influence measures satisfying those axioms by a formula. However, those properties are neither discussed with respect to indirect influence; nor are they in a fairness context.

Works on *indirect influence* can be broadly categorized into two groups: 1) law papers that discuss indirect influence/proxy discrimination in social and legal context; 2) methods that characterize indirect influence computationally. [12] to [14] lists some representative works in 1). Computational approaches are (almost) exhausted by [2] to [11] (where [10] and [11] are based on causality). Some prevailing indirect influence measures we focus on in this paper are summarized below.

- Orthogonal Feature Projection [5] and Black Box Auditing (BBA) [4] define indirect influence to be the change in the performance of a model when the features are altered to be less related to the protected attribute. While Orthogonal Feature Projection removes the linear relationship between the features and the protected attribute; BBA aims for pairwise independence between each feature and the protected attribute. Orthogonal Feature Projection and BBA require only black box access to the model. They are also at risk of generating out-of-distribution samples when they remove protected information from the features.
- Disentangling Influence [2] views data as generated from two independent factors of variation that represent protected and unprotected information. Similar to TCAV, Disentangling Influence computes indirect influence scores for individual instances. However, Disentangling

> Influence requires a high-quality disentangled representation, and in general the disentanglement is not unique.

We provide a further assessment of these methods in section 5 based on our proposed axioms.

3. Axioms of Indirect Influence

Given a learning model and prediction Y, let $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ be a function over a universal set of random variables (or random vectors). It represents a indirect influence measure (in $\mathbb{R}$) with respect to Y. The set of axioms for $\mathcal{II}_Y$ are defined below. We view them not as an exhaustive account of the properties of indirect influence, but as an incomplete set of desirable properties which we expect to evolve and grow.

Definition 3.1 (Universality)**.** A function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is universal if $\mathcal{II}_Y(Z_i)$ is *uniquely defined* for every random variable Z_i jointly distributed with Y.

Universality reflects that indirect influence is not constrained to variables within the model. While direct influence scores can only be computed for variables in the model, $\mathcal{II}_Y(Z_i)$ should be defined for all Z_i that are jointly distributed with Y. Further, an indirect influence measure should be trustworthy in the sense that for each $Z \in \mathcal{Z}$, it does not report multiple/conflicting values (e.g. due to a change in random seed, etc.). Otherwise, not only is such an indirect influence report hardly interpretable, it is also vulnerable to manipulation by users who prefers a certain result. Universality is not necessarily an axiom of indirect influence; rather, it is an inherent property of the function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ being well-defined.

Definition 3.2 (Symmetry)**.** A function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is *symmetric* if for any random variables Z_1 and Z_2 satisfying

$$(Z_1 \perp Y)|Z_2,\ (Z_2 \perp Y)|Z_1,$$

we have that $\mathcal{II}_Y(Z_1) = \mathcal{II}_Y(Z_2)$.

Symmetry states that if two sets of variables contain exactly the same information about Y, then they should have the same indirect influence (score). One consequence of symmetry is that if two random variables are identical in realization, i.e. $Z_i = Z_j$, then they have the same indirect influence.

Definition 3.3 (Null Variable)**.** A function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ satisfies the *null variable* criterion if for any random variable $Z \in \mathcal{Z}$ such that $S \perp Y$ we have that $\mathcal{II}_Y(Z) = 0$.

The Null Variable Criterion states that variables not correlated with predictions should have no indirect influence.

Definition 3.4 (Invariance)**.** We call a function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ *invariant* if for any random variable $Z \in \mathcal{Z}$ and any injective transformation f of Z, we have:

$$\mathcal{II}_Y(Z) = \mathcal{II}_Y(f(Z)).$$

Indirect influence measures are invariant to invertible transformations. For example, Z, $2Z$, and $Z+1$ should all have the same indirect influence on Y since each of these variables can be recovered from one another.

Definition 3.5 (Efficiency)**.** A function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is efficient if there exists a value $k \in \mathbb{R}$ such that for any $Z = (Z_1, \dots, Z_m)$ satisfying

(1) $I(S;Y) = 0$ for all subsets S of random variables in Z and $|S| \geq 2$ (e.g. $I(Z_1;Z_3;Y)$ with $S = \{Z_1, Z_3\}$);
(2) $H(Y|Z) = 0$,

we have that $\mathcal{II}_Y(Z) = \sum_{i=1}^m \mathcal{II}_Y(Z_i) = k$.

Note that $H(Y|Z)$ is the conditional entropy of Y given Z and $I(S;Y)$ is the interaction information (or co-information) of variables in S and Y – this is a generalization of mutual information to collections of more than two random variables. The efficiency criterion assigns meaning to the scale of an indirect influence function. It states that, for any set of variables (random vector) that decompose Y into independent factors, their indirect influence (as a set of variables or random vector) should be equal. While this may appear abstract at a first glance, Efficiency is indeed the product of two motivations. First, for any random variable that fully determines Y, they should have the same and maximum indirect influence on Y.

Definition 3.6 (Consistency)**.** A function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is *consistent* if there exists a $k \in \mathbb{R}$ such that for all random variable $Z \in \mathcal{Z}$ with $H(Y|Z) = 0$, we have $\mathcal{II}_Y(Z) = k$ and for all random variable $Z' \in \mathcal{Z}$, $\mathcal{II}_Y(Z') \leq k$. We call k the full influence of $\mathcal{II}_Y$.

Consistency provides a scale to which we can compare and measure the magnitude of $\mathcal{II}_Y(Z)$ for any variable Z. However, it is common that the variable Z fully determining Y is a joint random variable (i.e. $Z = (Z_1, Z_2, \dots, Z_n)$), whose indirect influence is hard to compute due to the overlapping information of Z_i about Y. This leads to another motivation behind Efficiency as follows

Definition 3.7 (Additivity)**.** A function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is *additive* if for all random variables $Z = (Z_1, \dots, Z_n)$ satisfying

$$I(S;Y) = 0 \text{ for all subsets } S \text{ of random variables in } Z,$$

we have $\mathcal{II}_Y(Z) = \sum_{i=1}^n \mathcal{II}_Y(Z_i)$

That is, if the components Z_i do not contain any redundant information about Y, the indirect influence of $Z = (Z_1, \dots, Z_n)$ is the sum of those of Z_i.[1] Additivity makes an indirect influence more interpretable and provides a way to compute its full influence – disentangling a variable Z that fully determines Y and adding up the indirect influence of the latent codes. We see that Consistency and Additivity are exactly what Efficiency is intended for by definition.

Corollary 3.8. *If a function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is consistent and additive, it is efficient.* □

4. Consequences of the Axioms

Given the proposed axioms in section 3, we are interested in if there exists a minimal set of properties such that they can derive all others. Further, can we characterize the family of functions that satisfy those axioms? We answer both questions by showing that an *efficient* function satisfies all the axioms (except Universality, which is considered as an inherent property of the function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ being well-defined rather than an axiom).

As preliminaries, we assume without loss of generality that variables in our work are all discrete and follow [15] (Chapter 2) on their definitions of the concepts in information theory. We also follow [16] and [17] on their definition of intersection information. Given a prediction Y, suppose $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is an indirect influence function, where $\mathcal{Z}$ is the set of random variables that are jointly distributed with Y.

We start by proving that Efficiency implies the Null Variable Criterion and Symmetry.

Proposition 4.1. *If a function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is efficient, it satisfies the null variable criterion.*

Proof. First, we point out that:

Lemma 4.2. *For any prediction Y, there exists a random variable $Z \in \mathcal{Z}$ such that $H(Y|Z) = 0$.*

Proof. Note that this lemma fails if we allow continuous variables (a uniform distribution of $\frac{1}{8}$ over $[0, 8]$) and indeed, that leads to a counterexample to the

[1] The rationale is that: 1) Z must contain all the indirect influence of Z_i as their ”parent”; 2) any addition/reduction in the indirect influence of Z (compared to the total indirect influence of Z_i) can only be justified by additional/redundant influence produced by putting Z_i together. Since there is no such influence (i.e. there is no statistical correlation between the overlaps of $Z_i, \dots, Z_j$ and Y), the indirect influence of Z is precisely the sum of those of Z_i.

proposition above[2]. Otherwise, for $Y \in \mathcal{Z}$, note that $H(Y|Y) = H(Y,Y) - H(Y) = H(Y) - H(Y) = 0$ as desired. □

Let $Z \in \mathcal{Z}$ be a variable that is independent of Y. By Lemma 4.2, note that $Y \in \mathcal{Z}$ itself satisfies $H(Y|Y) = 0$. We claim that $H(Y|(Z,Y)) = 0$ and the intersection information $I(Z;Y;Y) = 0$.

First, note that $Z \perp Y$ implies $Z \perp Y|Y$. By Corollary 2.92 [15], it follows that $I(Z;Y|Y) = 0$. Then, by Definition 2.60 [15], we have

$$H(Y|(Z,Y)) = H(Y|Y) - I(Z;Y|Y) = H(Y|Y) = 0,$$

as desired.

At the same time, we have $I(Z;Y;Y) = I(Z;Y) - I(Z;Y|Y) = 0 - 0 = 0$ [17]. Thus, since $\mathcal{II}_Y$ is efficient, it follows that

$$\mathcal{II}_Y(Y) + \mathcal{II}_Y(Z) = k,$$

with some constant $k \in \mathbb{R}$. Further, as $H(Y|Y) = 0$, we know from Efficiency again that $\mathcal{II}_Y(Y) = k$. Therefore, we conclude that $\mathcal{II}_Y(Z) = 0$ and $\mathcal{II}_Y$ satisfies null variable criterion as desired. □

Proposition 4.3. *If a function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is efficient, it is symmetric.*

Proof. We start by proving two lemmas.

Lemma 4.4. *For any random variable Z_1 and for any real number $0 \leq k \leq H(Z_1)$, there exists a random variable Z such that $I(Z_1;Z) = k$.*

Proof. Given a random variable Z_1 and a real number k, recall from definition of mutual information that for any random variable Z, we have

$$I(Z_1;Z) = \sum_{z_1,z} P(z_1,z) \log \frac{P(z_1,z)}{P(z_1)P(z)}.$$

In particular, we know from Theorem 2.4.1 and Corollary 2.90 [15] that $0 \leq I(Z_1;Z) \leq H(Z_1)$ where $I(Z_1;Z) = 0$ if and only if $Z_1 \perp Z$; and $I(Z_1;Z) = H(Z_1)$ if $Z_1 = Z$ (their joint distribution is the same as that of Z).[3]

Assume that Z and Z_1 have the same range with $|Z| = |Z_1| = n$ and that Z has the same probability measure as Z_1 (i.e. $P(z) = P(z_1)$ for $z_1 = z$, but

[2] Let Y be a uniform distribution of $\frac{1}{8}$ over $[0,8]$ and consider the function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ defined by $\mathcal{II}_Y(Z) = 42$ for all $Z \in \mathcal{Z}$. Since $H(Y|Z) \leq h(Z) < 0$ for all $Y \in \mathcal{Z}$, it vacuously satisfies Efficiency. At the same time, note that $\mathcal{II}(Z) = 42 \neq 0$ for any $Z \perp Y$, which violates the Null Variable Criterion.

[3] Again, this does not hold if continuous variables are under consideration with $I(Z_1;Z_1)$ undefined. However, I do believe the proposition holds in general by a similar argument.

their joint distribution $P(z_1, z)$ can be arbitrary). Then, we may consider $I(Z_1; Z)$ as a *continuous* function of the set of n^2-dimensional vectors

$$V = \{(P(z_1, z)) \mid P(z_1, z) \geq 0, \sum_{z_1} P(z_1, z) = P(z), \sum_{z} P(z_1, z) = P(z_1)\}.$$

Then, we already know from above that $I(Z_1; Z) = 0$ if $(P(z_1, z)) = (P(z_1)P(z))$ and $I(Z_1; Z) = H(Z_1)$ if $P(z_1, z) = P(z_1)$ for $z_1 = z$ and $P(z_1, z) = 0$ otherwise.

Now, suppose $(P(z_1, z)) = (P(z_1)P(z))$. Consider for some fixed z_1 and $z \neq z_1$, $P'(z_1, z) = P(z_1, z) - \epsilon$, $P'(z_1, z_1) = P(z_1, z_1) + \epsilon$, $P'(z, z_1) = P(z, z_1) - \epsilon$, $P'(z, z) = P(z, z) + \epsilon$, where $0 < \epsilon < P(z_1, z) = P(z, z_1)$. Let $(P'(z_1, z))$ be the same n^2-dimensional vector as $(P(z_1, z))$ except that the four probabilities above are replaced. Then, note that $(P'(z_1, z))$ is in V and let Z' be the same random variable as Z except that its joint distribution with Z_1 follows $(P'(z_1, z))$. By the derivative of the function $x \log \frac{x}{c}$ with $0 < c \leq 1$, note that

$$I(Z_1; Z') > I(Z_1; Z).$$

Now, we can apply the steps above to Z' and yield the same result, until $P(z_1, z) = P(z, z_1) = 0$ for all $z \neq z_1$ and $P(z_1, z_1) = P(z_1)$. Then, we switch to another z_1 and yield the same increasing pattern, until $P(z_1, z) = P(z_1)$ for all $z = z_1$ and $P(z_1, z) = 0$ otherwise. In other words, $I(Z_1; Z)$ is a continuously increasing function (in the direction of changes above) that ranges from 0 to $H(Z_1)$. This completes the proof. □

Lemma 4.5. *For any random variables $Z_1, Z_2 \in \mathcal{Z}$, there exists a random variable Z such that $H(Y|(Z_1, Z)) = 0$ and $I(Z_1; Z; Y) = I(Z_2; Z; Y) = 0$.*

Proof. Given two random variables $Z_1, Z_2 \in \mathcal{Z}$, let $k = H(Y) - I(Z; Y)$. Since $I(Z; Y) \leq H(Y)$, it follows that $0 \leq k \leq H(Y)$ and by Lemma 4.4, we can find a random variable $Z \in \mathcal{Z}$ such that $I(Z; Y) = k$. In particular, since $I(Z; Y)$ only depends on the distribution of Z, Y, and (Z, Y), we can specify the distributions of (Z, Z_1), (Z, Z_2), (Z, Z_1, Y), and (Z, Z_2, Y) with $I(Z; Y)$ unchanged. In particular, let

$$\begin{aligned} P(z, z_1) &= P(z)P(z_1), \\ P(z, z_2) &= P(z)P(z_2), \\ P(z|z_1, y) &= P(z|y), \\ P(z|z_2, y) &= P(z|y). \end{aligned}$$

Then, it follows that

$$\begin{aligned} P(z, z_1|y) &= P(z|z_1, y)P(z_1|y) = P(z|y)P(z_1|y), \\ P(z, z_2|y) &= P(z|z_2, y)P(z_2|y) = P(z|y)P(z_2|y). \end{aligned}$$

In other words, Z is not only independent of Z_1 and Z_2, but also conditionally independent of either of them given Y. This implies that

$$I(Z_1;Z;Y) = I(Z_1;Z) - I(Z_1;Z|Y) = 0 - 0 = 0,$$
$$I(Z_2;Z;Y) = I(Z_2;Z) - I(Z_2;Z|Y) = 0 - 0 = 0.$$

What remains is to verify that $H(Y|(Z_1,Z)) = 0$. To do this, since $I(Z_1;Z;Y) = I(Z_1;Y) - I(Z_1;Y|Z) = I(Z_2;Y) - I(Z_2;Y|Z)$, note that

$$I(Z_1;Y) = I(Z_1;Y|Z) \text{ and } I(Z_2;Y) = I(Z_2;Y|Z). \tag{4.1}$$

Therefore, we have

$$I(Y;(Z_1,Z)) = I(Z_1;Y|Z) + I(Z;Y) = I(Z_1;Y) + I(Z;Y) = H(Y).$$

Hence, it yields that $H(Y|(Z_1;Z)) = H(Y) - I(Y;(Z_1,Z)) = H(Y) - H(Y) = 0$ and the proof is complete. □

Now, given two random variables $Z_1, Z_2 \in \mathcal{Z}$ such that

$$Z_1 \perp Y|Z_2, \text{ and } Z_2 \perp Y|Z_2,$$

by Lemma 4.5, we can find a random variable $Z \in \mathcal{Z}$ such that $H(Y|(Z_1,Z)) = 0$ and $I(Z_1;Z;Y) = I(Z_2;Z;Y) = 0$. Then, we know from Proposition 6.2 that

$$I(Z_1;Y) = I(Z_2;Y).$$

Hence, by Equation 4.1, it follows that

$$I(Y;(Z_2,Z)) = I(Z_2;Y|Z) + I(Z;Y) = I(Z_2;Y) + I(Z;Y) = I(Z_1;Y) + I(Z;Y) = H(Y).$$

In other words, we have

$$H(Y|(Z_2,Z)) = H(Y) - I(Y;(Z_2,Z)) = H(Y) - H(Y) = 0.$$

Since $H(Y|(Z_1;Z)) = 0$, $I(Z_2;Z;Y) = I(Z_1;Z;Y) = 0$ and $\mathcal{II}_Y$ is efficient, we see that

$$\mathcal{II}_Y(Z) + \mathcal{II}_Y(Z_1) = \mathcal{II}_Y(Z) + \mathcal{II}_Y(Z_2) = k,$$

for some constant $k \in \mathbb{R}$. Therefore, we conclude that $\mathcal{II}_Y(Z_1) = \mathcal{II}_Y(Z_2)$ and $\mathcal{II}_Y$ is symmetric as desired. □

To show that Efficiency implies Invariance, it suffices to verify that Symmetry implies Invarience by Proposition 4.3.

Proposition 4.6. *If a function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is symmetric, it is invariant.*

Proof. Given a random variable $Z \in \mathcal{Z}$ and an injective function $f : Z \to \overline{Z}$, note that f is bijective when restricted to $f : Z \to f(Z)$ with $f^{-1} : f(Z) \to Z$

its inverse. Let $Z' = f(Z)$. By Symmetry, we want to show that $Z \perp Y|Z'$ and $Z' \perp Y|Z$. First, it follows from definition of conditional probability that

$$P(Z = z|Z' = z') = \begin{cases} 1 & z = f^{-1}(z'), \\ 0 & z \neq f^{-1}(z'). \end{cases} \text{ for any } z' \in Z', \tag{4.2}$$

$$P(Z' = z'|Z = z) = \begin{cases} 1 & z' = f(z), \\ 0 & z' \neq f(z). \end{cases} \text{ for any } z \in Z, \tag{4.3}$$

Similarly, we have

$$P(Z = z|Y = y, Z' = z') = \begin{cases} 1 & z = f^{-1}(z'), \\ 0 & z \neq f^{-1}(z'). \end{cases} \text{ for any } y \in Y, z' \in Z',$$

$$P(Z' = z'|Y = y, Z = z) = \begin{cases} 1 & z' = f(z), \\ 0 & z' \neq f(z). \end{cases} \text{ for any } y \in Y, z \in Z.$$

By Bayes Theorem for conditional joint distribution, this implies that

$$\begin{aligned} P(Z = f^{-1}(z'), Y = y|Z' = z') &= P(Z = f^{-1}(z')|Y = y, Z' = z')P(Y = y|Z' = z') \\ &= P(Y = y|Z' = z'), \end{aligned}$$

and in general, it yields that

$$P(Z = z, Y = y|Z' = z') = \begin{cases} P(Y = y|Z' = z') & z = f^{-1}(z'), \\ 0 & z \neq f^{-1}(z'). \end{cases} \text{ for any } y \in Y, z' \in Z', \tag{4.4}$$

$$P(Z' = z', Y = y|Z = z) = \begin{cases} P(Y = y|Z = z) & z' = f(z), \\ 0 & z' \neq f(z). \end{cases} \text{ for any } y \in Y, z \in Z, \tag{4.5}$$

Hence, we see from (2.2) and (2.4) that for any $z \in Z$, $z' \in Z'$, and $y \in Y$, it satisfies that

$$\begin{aligned} P(Z = z|Z' = z')P(Y = y|Z' = z') &= \begin{cases} 1 \cdot P(Y = y|Z' = z') = P(Y = y|Z' = z') & z = f^{-1}(z'), \\ 0 \cdot P(Y = y|Z' = z') = 0 & z \neq f^{-1}(z'). \end{cases} \\ &= P(Z = z, Y = y|Z' = z'). \end{aligned}$$

That is, $Z \perp Y|Z'$. Similarly, we have

$$\begin{aligned} P(Z' = z'|Z = z)P(Y = y|Z = z) &= \begin{cases} 1 \cdot P(Y = y|Z = z) = P(Y = y|Z = z) & z = f(z), \\ 0 \cdot P(Y = y|Z = z) = 0 & z \neq f(z). \end{cases} \\ &= P(Z' = z', Y = y|Z = z), \end{aligned}$$

and $Z' \perp Y|Z$. Hence, since $Z' = f(Z)$ and $\mathcal{II}_Y$ is symmetric, we conclude that

$$\mathcal{II}_Y(Z) = \mathcal{II}_Y(f(Z)),$$

and $\mathcal{II}_Y$ is invariant as desired. □

Hence, by Proposition 4.3 and Proposition 4.6, we have

Corollary 4.7. *If a function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ is efficient, it is invariant.* □

Note that the converse of Proposition 4.3 is not true.

Proposition 4.8. *Symmetry does* not *imply Null Variable Criterion.*

Proof. Consider the function $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ defined by $\mathcal{II}_Y(Z) = I(Z;Y) + 42$. We know from Proposition 6.2 (later) that $I(Z;Y)$ is symmetric. This implies that $\mathcal{II}_Y$ is symmetric as well (it is merely a translation of $I(Z;Y)$). At the same time, since $\mathcal{II}_Y(Z) = I(Z;Y) + 42 = 0 + 42 = 42 \neq 0$ for any variable $Z \perp Y$, we conclude that $\mathcal{II}_Y(Z)$ does not satisfy the null variable criterion. □

Corollary 4.9. *Symmetry does* not *imply Efficiency.* □

In summary, it follows from Proposition 4.1, 4.3 and Corollary 4.7 that efficient functions (that are well-defined over $\mathcal{Z}$) satisfy all proposed axioms.

5. Analysis of Existing Indirect Influence Methods

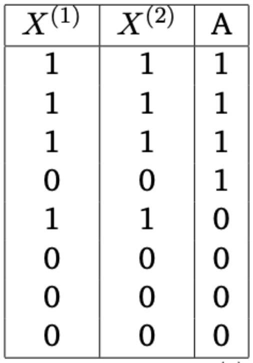
As an application, we discuss possible issues of existing indirect influence methods (namely, BBA, Disentangling Influence, and Orthogonal Feature Projection) with respect to the proposed axioms. A method that fails to satisfy some of those natural properties may be poorly aligned with part of goals of indirect influence. On the other hand, it is also important to note that operationalizing indirect influence measure may require sacrificing desirable properties in the name of computational feasibility. Hence, we pose these critiques primarily to document trade-offs which have been made and highlight space for potential improvements.

5.1. **Black Box Auditing (BBA).**

We provide an example to show that BBA is not *universal* in the sense that it may assign multiple values to the indirect influence of a variable.

Consider an audit for the dataset in Figure 2

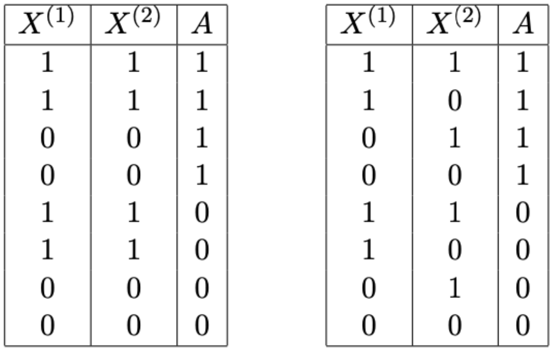
The BBA algorithm perturbs feature $X^{(1)}$ so that the conditional distributions for $X^{(1)}|A = 0$ and $X^{(1)}|A = 1$ are the same. Specifically, both conditional distributions are mapped to the median distribution while minimizing the number of perturbations. There are 9 possible ways to optimally



$X^{(1)}$	$X^{(2)}$	A
1	1	1
1	1	1
1	1	1
0	0	1
1	1	0
0	0	0
0	0	0
0	0	0

FIGURE 2. Toy dataset with unprotected attributes $X^{(1)}$, $X^{(2)}$ and sensitive attribute A.

alter $X^{(1)}$ to these ends, and similarly for $X^{(2)}$. We show two of the $9 \times 9 = 81$ such solutions below.



$X^{(1)}$	$X^{(2)}$	A
1	1	1
1	1	1
0	0	1
0	0	1
1	1	0
1	1	0
0	0	0
0	0	0

$X^{(1)}$	$X^{(2)}$	A
1	1	1
1	0	1
0	1	1
0	0	1
1	1	0
1	0	0
0	1	0
0	0	0

FIGURE 3. Two optimal solutions to the feature repairing task in BBA.

We see that in one repair solution the random vector $(X^{(1)}, X^{(2)})$ is restricted to $(0,0)$ and $(1,1)$ whereas in the other solution, $(X^{(1)}, X^{(2)})$ is uniform over $(0,0), (0,1), (1,0)$, and $(1,1)$. To compute the indirect influence of $X^{(1)}$, the BBA algorithm compares the predictions of the learning model on the original and perturbed dataset. In particular, models which are sensitive to the distribution of $(X^{(1)}, X^{(2)})$ (e.g. $M(x(1), x(2)) = x(1) + x(2)$) would result in different indirect influence reports on the two repairs and the method is not universal.

5.2. **Disentangling Influence.** Again, we provide an example to show that Disentangling Influence is not *universal* in the sense that it may assign multiple values to the indirect influence of a variable.

Consider a dataset $X = \{x_i\}_{i=1}^n$ with instances $x_i = (a_i, b_i, c_i)$ where a_i, b_i, c_i are i.i.d. Unif(0, 1). Let $y_i = a_i + b_i$ be the label of instance i. The disentangling influence method finds a disentangled representation for X. That is, a representation $Z = (Z_1, Z_2, Z_3)$ for A, B, C such that Z_1, Z_2, Z_3 are independent and A, B, C can be reconstructed from Z. We denote the reconstructions of A, B, and C by $\hat{A} = \hat{A}(Z)$, $\hat{B} = \hat{B}(Z)$, and $\hat{C} = \hat{C}(Z)$.

Let $Z = (Z_1, Z_2, Z_3)$ be defined as follows:

$$\begin{aligned} Z_1 &= (A - C) \mod 1, \\ Z_2 &= (B - C) \mod 1, \\ Z_3 &= C. \end{aligned}$$

We claim that

Claim 5.1. *Z_1, Z_2, Z_3 are mutually independent.*

Proof. We start by showing that Z_1 and Z_2 are independent. Consider Z_1 and Z_2 as points on a circle of circumference 1 by wrapping the unit interval so that zero lies at the top of the circle. To compute Z_1, starting at the point on the circle that corresponds to the value of A, we travel around the circle a distance of C. Similarly, Z_2 is computed by starting at B instead of A. Before subtracting C, Z_1 and Z_2 are independently and uniformly distributed around the circle. Hence, rotating both by a distance of C preserves this independence. In other words, Z_1 and Z_2 are independent.

Next, we prove that Z_1 and Z_3 are independent. To generate Z_1, we start somewhere around the circle with uniformity and rotate by C. Similarly, the rotation preserves the uniform distribution around the circle for Z_1 and it follows that Z_1 and C are independent. Recalling that $Z_3 = C$, we have Z_1 and Z_3 are independent. By the same argument, we know that Z_2 and Z_3 are independent.

At this point, we have verified that Z_1, Z_2, Z_3 are pairwise independent. What remains is to show that each pair of variables Z_i, Z_j are independent conditioned on the third variable, Z_k. We first show that Z_1 and Z_2 are independent given Z_3. Since A, B are independent and uniformly distributed around the circle, knowing both Z_2 (resp. Z_1) and Z_3 preserves a uniform distribution for Z_1 (resp. Z_2) around the circle. In other words, given $Z_3 = C$, how far we rotate in the computation of Z_1 and Z_2 does not change the distribution of Z_1 and Z_2. Therefore, Z_1 and Z_2 are independent conditioned on Z_3. Lastly, we show that Z_1 and Z_3 are independent conditioned on Z_2. In other words, learning the final location of Z_1 (resp. Z_3) and Z_2 tells us nothing about how far the two were rotated. This is true due to the uniformity of A and B, so we see that Z_1 and Z_3 are independent conditioned on Z_2. By the same argument, we also know that Z_2 and Z_3 are independent given Z_1.

Therefore, we conclude that Z_1, Z_2, Z_3 are mutually independent. □

Now, consider the variables

$$\begin{aligned}\hat{A} &= (Z_1 + Z_3) \mod 1,\\ \hat{B} &= (Z_2 + Z_3) \mod 1,\\ \hat{C} &= Z_3.\end{aligned}$$

We see by definition of Z_1, Z_2, Z_3 that $\hat{A}, \hat{B}$ and $\hat{C}$ are reconstruction of A, B, C from Z_1, Z_2, Z_3. Thus, $Z = (Z_1, Z_2, Z_3)$ is a valid disentangled representation for X. Note, however, that $Z' = (A, B, C)$ is also a valid disentangled representation for X. Using these two different representations, we can get distinct indirect influence reports by selecting models that are sensitive to the difference. Hence, the disentangling influence method is not universal and the learned disentanglement can have significant effect on the result.

5.3. **Orthogonal Feature Projection.** Orthogonal Feature Projection method is not (always) invariant under an injective transformation of sensitive variable A. For example, consider the case where $A \sim \text{Unif}(1)$, $X = A$, and a model $M(x) = x$ for $x \in X$ (where the labels are 0 and 1). Then, it follows that $X_\perp = 1 = X$ and $M(X_\perp) - M(X) = 0$. In other words, A is assigned a zero indirect influence. Now, consider the transformed variable $A - 1$. Then, we have $X_\perp = 0$ and $M(X_\perp) - M(X) = -1$. Hence, an indirect influence of -1 is assigned to $A - 1$. While the decrement operation is injective, Orthogonal Feature Projection yields significantly different indirect influence scores to A and $A - 1$ for some model M. As a result, the method does not satisfy the Invariance criterion and it is possible to game Orthogonal Feature Projection by accessing A under some transformation.

6. Mutual Information

While the four existing indirect influence measures all break some axioms, we claim that $I_Y : \mathcal{Z} \to \mathbb{R}$ defined by $I_Y(Z) = I(Z, Y)$, i.e. the mutual information between a given variable $Z \in \mathcal{Z}$ and Y, has all desired properties. As we have proved in section 4, it suffices to show that I_Y satisfies Universality and Efficiency.

Proposition 6.1. *$I_Y : \mathcal{Z} \to \mathbb{R}$ is universal.*

Proof. Given a random variable $Z \in \mathcal{Z}$ that is jointly distributed with Y, we know by definition of mutual information that $I(Z, Y)$ is well-defined and it is universal. □

Proposition 6.2. *$I_Y : \mathcal{Z} \to \mathbb{R}$ is efficient.*

Proof. In particular, we claim that for any random variable $Z = (Z_1, Z_2, \ldots, Z_n)$ that satisfies the assumptions in Efficiency, we have $I_Y(Z) = H(Y)$. To see why, we start with some intuitions.

First, the relationship among entropy, mutual information and intersection information (of three variables) can be represented with Figure 4.

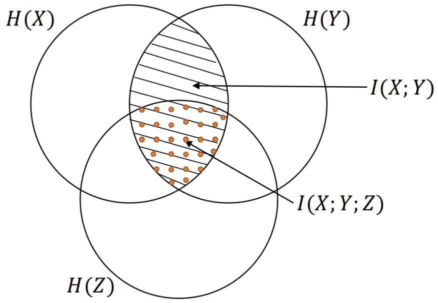


FIGURE 4. Graphical representation of information metrics [17].

When $H(Y|(X,Z)) = 0$, we know that $I(Y;(X,Z)) = H(Y)$. That is, the circle of $H(Y)$ is contained in the union of $H(X)$ and $H(Z)$. At the same time, since $I(X;Y;Z) = 0$, we know that $H(Y)$ does not intersect with the intersection of $H(X)$ and $H(Y)$. As a result, we have Figure 5 and it demonstrates that $I(X;Y) + I(Z;Y) = I((X;Z),Y) = H(Y)$.

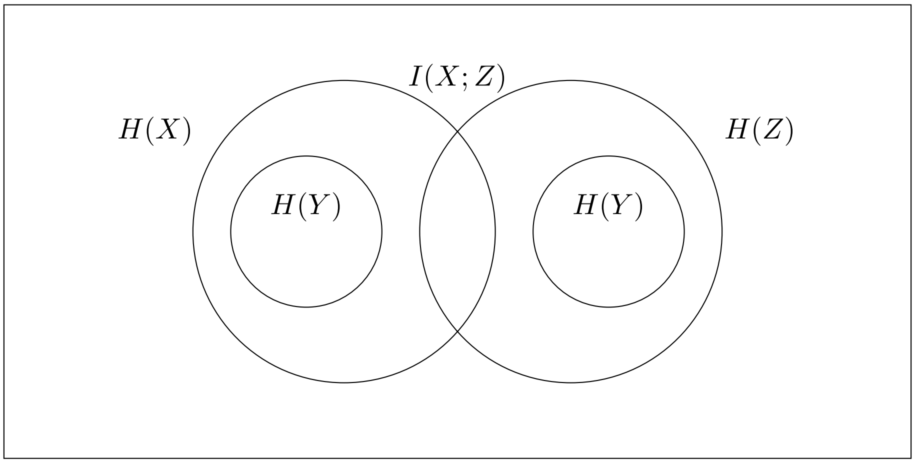


FIGURE 5. Graphical representation of information metrics under the assumptions of Efficiency.

Mathematically, for any random variable $Z = (Z_1, Z_2)$ with $I(Z_1; Z_2; Y) = 0$ and $H(Y|Z) = 0$, we know that

$$I_Y(Z_1) + I_Y(Z_2) = H(Z_1) + H(Y) - H(Z_1, Y) + H(Z_2) + H(Y) - H(Z_2, Y).$$

By definition of intersection information, it follows that

$$\begin{aligned} I_Y(Z_1) + I_Y(Z_2) &= H(Y) + H(Z_1, Z_2) - H(Z_1, Z_2, Y). + I(Z_1; Z_2; Y) \\ &= H(Y) + H(Z_1, Z_2) - H(Z_1, Z_2, Y). \end{aligned}$$

Now, since $H(Y|(Z_1, Z_2)) = 0$ and $H(Z_1, Z_2, Y) = H(Z_1, Z_2) + H(Y|(Z_1, Z_2))$ (chain rule), we have $H(Z_1, Z_2) = H(Z_1, Z_2, Y)$ and $I_Y(Z_1) + I_Y(Z_2) = H(Y)$ as desired.

In fact, the intuition given by the Venn graph converts the problem into a problem of Inclusion-exclusion Principle. That said, one can apply the principle to verify the following result in general.

Proposition 6.3. *For any random variable $Z = (Z_1, Z_2, \dots, Z_n)$,*

$$\sum_{i=1}^{n} I_Y(Z_i) = \sum_{S} (-1)^{|S|} I(S; Y) + H(Y) + H(Z_1, Z_2, \dots, Z_n) \\ - H(Z_1, Z_2, \dots, Z_n, Y),$$

where S denotes any subset of random variables in Z with $|S| \geq 2$.

Then, if Z satisfies the assumptions of Efficiency, it follows immediately from the proposition above that $\sum_{i=1}^{n} I_Y(Z_i) = H(Y)$.

We can also prove more formally an alternate version of the proposition above that avoids unnecessary technicalities.

Lemma 6.4. *For any random variable $Z = (Z_1, Z_2, \dots, Z_n)$ with $I(S; Y) = 0$, where S is any subset of random variables in Z such that $|S| \geq 2$, it satisfies that*

$$\sum_{i=1}^{n} I_Y(Z_i) = H(Y) + H(Z_1, Z_2, \dots, Z_n) - H(Z_1, Z_2, \dots, Z_n, Y).$$

Proof. The result follows immediately from Proposition 6.3. Otherwise, we may prove it by induction.

Base case ($n = 1$). In this case, we know directly from [15] that

$$I_Y(Z_1) = I(Z_1, Y) = H(Y) + H(Z_1) - H(Z_1, Y),$$

which satisfies the lemma above.

Inductive step. Assume the statement holds for $n = k$ ($k \geq 1$), we want to show that it holds for $n = k + 1$.

First, we know by inductive hypothesis that

$$\begin{aligned}\sum_{i=1}^{k+1} I_Y(Z_i) &= I_Y(Z_{k+1}) + \sum_{i=1}^{k} I_Y(Z_i)\\ &= H(Y) + H(Z_{k+1}) - H(Z_{k+1}, Y)\\ &+ H(Y) + H(Z_1, Z_2, \ldots, Z_k) - H(Z_1, Z_2, \ldots, Z_k, Y),\end{aligned}$$

where S denotes any subset of random variables in $Z - Z_{k+1}$ with $|S| \geq 2$. Now, consider $(Z_1, Z_2, \ldots, Z_k)$ as a random variable (random vector) and Z_{k+1} as another, by the same argument as the three-variable case, we have

(6.1)

$$\begin{aligned}&H(Y)+H(Z_{k+1})-H(Z_{k+1},Y)+H(Y)+H(Z_1,Z_2,\ldots,Z_k)-H(Z_1,Z_2,\ldots,Z_k,Y)\\ &= I((Z_1,Z_2,\ldots,Z_k);Z_{k+1};Y)+H(Y)+H(Z_1,Z_2,\ldots,Z_{k+1})-H(Z_1,Z_2,\ldots,Z_{k+1},Y).\end{aligned}$$

Similarly, considering $(Z_1, Z_2, \ldots, Z_{k-1})$ as a random vector and Z_k as another, we know from definition of intersection information that

$$\begin{aligned}&I((Z_1,Z_2,\ldots,Z_k);Z_{k+1};Y)\\ &= I((Z_1,Z_2,\ldots,Z_{k-1});Z_{k+1};Y)+I(Z_k;Z_{k+1};Y)-I((Z_1,Z_2,\ldots,Z_{k-1});Z_k;Z_{k+1};Y).\end{aligned}$$

Proceeding inductively, we can decompose any intersection information of form $I(Z'; S; Z_{k+1}; Y)$ (where Z' is a random vector consisting of Z_i and S is a subset of random variables in $Z - Z_k - Z'$) until every term on the right hand side of the equality above is in the form $I(S'; Z_k; Y)$, where S' is a subset of random variables in $Z - Z_k$. By assumption, this implies that

$$I((Z_1, Z_2, \ldots, Z_k); Z_{k+1}; Y) = 0.$$

Hence, by Equation 6.1, we conclude that

$$\sum_{i=1}^{k+1} I_Y(Z_i) = H(Y) + H(Z_1, Z_2, \ldots, Z_{k+1}) - H(Z_1, Z_2, \ldots, Z_{k+1}, Y),$$

as desired. □

Now, by Lemma 6.4 and the chain rule that $H(Y|Z) = 0$ implies $H(Z_1, Z_2, \ldots, Z_n) = H(Z_1, Z_2, \ldots, Z_n, Y)$, we have $\sum_{i=1}^{n} I_Y(Z_i)$ as required. □

By Propositions 4.1, 4.3, 4.6, and 6.2, we see that

Corollary 6.5. $I_Y : \mathcal{Z} \to \mathbb{R}$ *is symmetric, invariant, and satisfies null variable criterion.* □

7. Discussion

Indirect influence measure is an important task for improving interpretability and evaluating fairness in machine learning models. In order to formalize this notion and better understand the strengths and drawbacks of various indirect influence methods, we introduce a set of desirable properties for indirect influence and evaluate existing methods in terms of these properties. Further, it is worth noting that efficient functions (that are well-defined over the set of variables jointly distributed with predictions) satisfy all proposed axioms and in particular, the mutual information with prediction Y is such an indirect influence measure.

We acknowledge two arguments against *Efficiency* as an *axiom*. First, the use of information metrics such as conditional entropy and interaction information is controversial; it is reasonable to argue for any other multivariate information measures as replacement. Another problem is on the assumption of additivity/linearity of indirect influence. In general, Efficiency assumes that if we keep adding variables that do not contain extra or redundant information as previous ones, indirect influence should vary linearly. However, this seems to be a stereotype from operations in $\mathbb{R}$ and it can be arbitrarily flexible for how two indirect influences $\mathcal{II}_Y(Z_1)$ and $\mathcal{II}_Y(Z_2)$ interact, depending on the structure of the space and meaning of the specific $\mathcal{II}_Y$. For example, the canonical indirect influence obtained from predictability does not grow linearly in this way. Therefore, it may be more appropriate to consider Efficiency as a desirable *human property* (rather than an axiom that *requires* every indirect influence method to have). In particular, we have seen that Efficiency characterizes a family of functions that satisfy those natural axioms. Further, it ensures a better interpretability and computability of an indirect influence method.

For future work, we recognize that the proposed axioms may be too theoretical to be verified universally on the main stream of empirical indirect influence methods. While the probability distributions and information metrics can be easily checked on one instance of the measures; it is difficult to prove a general case without prior knowledge of the variables. Further, the axioms may be too strict for empirical approaches that allow some uncertainties. Under these considerations, one next step is to (re)formulate our proposed axioms that are more accessible empirically.

We are also interested in associating the axioms with the concerns and social needs of indirect influence discussed in legal context. For example, Prince et al. (2019) considers that it is important to factor out the indirect influence of a protective attribute A in that of a proxy X. To do this, we may want to measure the indirect influence of A (on prediction Y) through X. One way to formalizing this question is to adopt an axiomatic approach.

In particular, given an indirect influence $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$, we want to find a function $\overline{\mathcal{II}}_Y : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$ such that

(1) $\overline{\mathcal{II}}_Y$ is well-defined for any pair of variables (Z, X) where $Z \in \mathcal{Z}$ is a variable jointly distributed Y and $X \in \mathcal{X}$ is a variable in the model;
(2) For any pair of variable $(Z, X) \in \mathcal{Z} \times \mathcal{X}$, if $I(Z; X; Y) = 0$, then $\overline{\mathcal{II}}_Y(Z, X) = 0$ (i.e. if Z and X do not share any information about Y, then X is not a proxy of Z);
(3) For $Z \in \mathcal{Z}$ and $X_1, X_2 \in \mathcal{X}$, if $I(Z; X_1; Y) \geq I(Z; X_2; Y)$, then $\overline{\mathcal{II}}_Y(Z, X_1) \geq \overline{\mathcal{II}}_Y(Z, X_2)$ (i.e. if Z shares more information about Y with X_1 than X_2, then Z has more indirect influence on Y via X_1 than X_2);
(4) For $Z \in \mathcal{Z}$ and $X = (X_1, \ldots, X_n) \in \mathcal{X}$ such that $I(X; Z; Y) = I(Z, Y)$ and $I(\mathcal{S}; Z; Y) = 0$ for any subset $\mathcal{S}$ of variables in X; $\overline{\mathcal{II}}_Y(Z, X) = \sum_{i=1}^{n} \overline{\mathcal{II}}_Y(Z, X_i) = \mathcal{II}_Y(Z)$ (i.e. if X contains all the information that Z has about Y and for every components X_i, X_j, there is no redundancy of such information; then the indirect influence of Z on Y via X is precisely all its indirect influence and it can be decomposed componentwise).

In fact, if $\mathcal{II}_Y : \mathcal{Z} \to \mathbb{R}$ already satisfies our proposed axioms as an indirect influence measure, we claim that it suffices to check $\overline{\mathcal{II}}_Y$ satisfies property (4) (while other are inherited from $\mathcal{II}_Y$ and property (4)). For $\mathcal{II}_Y$ defined by $\mathcal{II}_Y(Z) = I(Z, Y)$ (i.e. the mutual information with Y), we claim that $\overline{\mathcal{II}}_Y : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$ defined by $\overline{\mathcal{II}}_Y(Z, X) = I(Z, X, Y)$ (i.e. the intersection information with Y) is the unique measure that satisfies the above properties.

References

[1] J. Sliwinski; M. Strobel; and Y. Zick. “Axiomatic Characterization of Data-Driven Influence Measures for Classification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017. URL: https://arxiv.org/abs/1708.02153.

[2] C. T. Marx; R. L. Phillips; S. A. Friedler; C. Scheidegger; and S. Venkatasubramanian. “Disentangling influence: Using disentangled representations to audit model predictions”. In: (2019). URL: https://arxiv.org/abs/1906.08652.

[3] 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. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2783311. URL: https://arxiv.org/abs/1412.3756.

[4] A. Philip; C. Falk; S. A. Friedler; T. Nix; G. Rybeck; C. Scheidegger; B. Smith; and S. Venkatasubramanian. “Auditing black-box models for indirect influence”. In: *Knowledge and Information Systems 54.1*. 2018, pp. 95–122. URL: https://arxiv.org/abs/1602.07043.

[5] A. Julius; and L. Kagal. “Iterative orthogonal feature projection for diagnosing bias in black-box models”. In: *Conference oon Fairness, Accountabiligy, and Transparency in Machine Learning*. 2016. URL: https://arxiv.org/abs/1611.04967.

[6] K. Been; M. Wattenberg; J. Gilmer; C. Cai; J. Wexler; F. Viegas; and R. Sayres. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors TCAV”. In: *ICML*. 2018. URL: https://arxiv.org/abs/1711.11279.

[7] P. W. Koh; and P. Liang. “Understanding Black-box Predictions via Influence Functions”. In: *Proceedings of the 34 th International Conference on Machine Learning*. 2017. URL: https://arxiv.org/abs/1703.04730.

[8] S. Yeom; A. Datta; and M. Fredrikson. “Hunting for Discriminatory Proxies in Linear Regression Models”. In: *32nd Conference on Neural Information Processing Systems*. 2018. URL: https://par.nsf.gov/servlets/purl/10095671.

[9] A. Datta; M. Fredrikson; G. Ko; P. Mardziel; and S. Sen. “Proxy Non-Discrimination in Data-Driven Systems”. In: 2017. URL: https://arxiv.org/abs/1707.08120.

[10] H. Farbmacher; M. Huber; L. Lafférs; H. Langen; M. Spindler. “Causal mediation analysis with double machine learning”. In: 2020. URL: https://arxiv.org/abs/2002.12710.

[11] N. Kilbertus; M. Rojas-Carulla; G. Parascandolo; M. Hardt; D. Janzing; B. Schölkopf. “Avoiding Discrimination through Causal Reasoning”. In: *31st Conference on Neural Information Processing Systems*. 2017. URL: https://arxiv.org/abs/1706.02744.

[12] A. Prince; and D. Schwarcz. “Proxy Discrimination in the Age of Artificial Intelligence and Big Data”. In: *105 Iowa Law Review 1257*. 2019. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3347959.

[13] M. Boyarskaya; S. Barocas; and H. Wallach. “What Is a Proxy and Why Is It a Problem?” In: *ICML 2020 Workshop on Law Machine Learning (LML)*. 2020. URL: https://sites.google.com/view/icml-law-and-ml-2020/abstract-m-boyarskaya.

[14] F. J. Z. Borgesius. “Strengthening legal protection against discrimination by algorithms and artificial intelligence”. In: *The International Journal of Human Rights, Vol 24, 2020, Issue 10*. 2019. URL: https://www.tandfonline.com/doi/full/10.1080/13642987.2020.1743976.

[15] T.M. Cover; and J.A. Thomas. *Elements of Information Theory*. John Wiley Sons, Inc., Hoboken, New Jersey., 2006. ISBN: 978-0-471-24195-9.

[16] A. Jakulin; and I. Bratko. “Quantifying and Visualizing Attribute Interactions: An Approach Based on Entropy”. In: (2004). URL: https://arxiv.org/abs/1701.08868.

[17] A. Ghassam; and N. Kiyavash. “Interaction Information for Causal Inference: The Case of Directed Triangle”. In: (2017). URL: https://arxiv.org/abs/1701.08868.