Why Fair Labels Can Yield Unfair Predictions: Graphical Conditions for Introduced Unfairness

Carolyn Ashurst,[1] **Ryan Carey,**[2] **Silvia Chiappa,**[3] **Tom Everitt**[3]

[1]Alan Turing Institute, [2]University of Oxford, [3]DeepMind
cashurst@turing.ac.uk, ry.duff@gmail.com, csilvia@deepmind.com, tomeveritt@deepmind.com

Abstract

In addition to reproducing discriminatory relationships in the training data, machine learning (ML) systems can also introduce or amplify discriminatory effects. We refer to this as *introduced unfairness*, and investigate the conditions under which it may arise. To this end, we propose *introduced total variation* as a measure of introduced unfairness, and establish graphical conditions under which it may be incentivised to occur. These criteria imply that adding the sensitive attribute as a feature removes the incentive for introduced variation under well-behaved loss functions. Additionally, taking a causal perspective, *introduced path-specific effects* shed light on the issue of when specific paths should be considered fair.

1 Introduction

It is often said that "unfair data leads to unfair models", because machine learning systems tend to learn biases present in the training data. However, sometimes a model can produce unfair predictions even when the training labels are fair. More generally, a model can amplify the unfairness present in training labels.

To quantify this effect, which we refer to as *introduced unfairness*, we propose computing a suitable measure of disparity for the training labels and the model predictions, and then comparing the two values.

One such measure is the *total variation* (Zhang and Bareinboim 2018b), a generalisation of demographic disparity that describes the strength of the statistical relationship between a sensitive variable (such as gender) and an outcome (such as the score given to an applicant's resume). If the total variation of the predictions is greater than that of the training labels we say that there is *introduced total variation* (§3).

Introduced total variation is distinct from existing measures of unfairness like *separation* and *sufficiency*, which generalise *equalised odds* and *predictive parity* respectively (§4). For binary classifiers, separation prevents introduced total variation, while sufficiency prevents reduced total variation. In contrast, absence of introduced total variation guarantees neither separation or sufficiency.

Introduced unfairness would seem to be avoidable, since a perfectly accurate predictor would avoid it. This raises the questions: "*When* is there introduced unfairness?" and "*How* can it be removed?" To answer these questions, we use structural causal models (SCMs) and their associated graphs to represent the relationships between the variables underlying the training data (§2). We also build on influence diagrams, by including the predictor and the loss function of the machine learning system in the same graph. This allows us to investigate the conditions for introduced total variation for optimal predictors (§5).

We show that the class of loss function influences the graphical conditions under which introduced total variation is incentivised (§5.2). In particular, we consider *P-admissible* loss functions (Miller, Goodman, and Smyth 1993) — those for which it is optimal to output the expected label given the input features — such as mean squared error and cross-entropy. If the loss is P-admissible, then a stricter graphical criterion holds, meaning that a change of loss function is sometimes enough to prevent introduced unfairness.

The notion of introduced unfairness can be applied to causal definitions of fairness as readily as statistical ones (§6). In particular, *path-specific effects* (Pearl 2001) can help with understanding and addressing complex unfairness scenarios that are relevant to many real-world applications (Kilbertus et al. 2017; Chiappa 2019; Nabi and Shpitser 2018). We define *path-specific introduced effects* as the difference in some path-specific effect on labels and predictions. Building on this measure, we present some new considerations regarding the use of path-specific effects in machine learning fairness.

In Section 7, we study how often introduced total variation arises, empirically. We then discuss related work (§8) and discuss findings, limitations, and how these results can be applied (§9).

2 Setup

Our fairness analysis focuses on supervised learning algorithms used to make predictions about individuals, specifically regression and classification, and uses structural causal models (SCMs) to represent relationships among variables. Note that Section 6 relies upon the causal nature of SCMs, whereas the results of Sections 3, 4, 5 could also be translated into Bayesian Networks (Pearl 1986; Koller and Friedman 2009).

Definition 1 (Structural causal model (SCM); Pearl 2009; Pearl, Glymour, and Jewell 2016)**.** *A* structural causal model *$\mathcal{M}$ is a tuple $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{\mathcal{E}}) \rangle$, where $\boldsymbol{\mathcal{E}}$ is a set of exogenous (unobserved or latent) variables and $\boldsymbol{V}$ is a set of endogenous (observed) variables. $\boldsymbol{F}$ is a set of deterministic functions $\boldsymbol{F} = \{f_V\}$, where f_V determines the value of $V \in \boldsymbol{V}$ based on endogenous variables $\boldsymbol{Pa}^V \subseteq \boldsymbol{V} \setminus \{V\}$ and exogenous variables $\boldsymbol{\mathcal{E}}^V \subseteq \boldsymbol{\mathcal{E}}$, that is, $V \leftarrow f_V(\boldsymbol{Pa}^V, \boldsymbol{\mathcal{E}}^V)$. $P(\boldsymbol{\mathcal{E}})$ is a joint distribution over the exogenous variables, which is assumed to factorise.*

An SCM $\mathcal{M}$ may be associated with a directed graph $\mathcal{G}$ which has a node for each variable B and an edge $A \to B$ for every $A \in \mathbf{Pa}^B \cup \boldsymbol{\mathcal{E}}^B$. Paths from V_1 to V_2 of arbitrary length are denoted $V_1 \dashrightarrow V_2$. We only consider SCMs for which the graph is acyclic, and thus refer to $\mathcal{G}$ as the *associated directed acyclic graph (DAG)*, or *associated graph*. We say that $\mathcal{M}$ is *compatible* with $\mathcal{G}$. We often omit the exogenous variables from the graphs.

We define an *SL SCM* to be an SCM containing endogenous variables $Y, \hat{Y}, U$, representing the outcome variable, its model prediction, and the loss, of some SL model, respectively. Specifically, the loss function f_U is real-valued and has two arguments Y and $\hat{Y}$ (we consider the mean squared error $f_U = -(Y - \hat{Y})^2$ and zero-one loss $f_U = 0$ if $Y = \hat{Y}$; -1 otherwise). The parents of $\hat{Y}$ represent the input features. It may be the case that inputs to $\hat{Y}$ are descendants of Y. The associated DAG is called an *SL graph*. An SL SCM (or graph) includes a *sensitive variable* A if it has an endogenous variable/node A, which represents a sensitive attribute such as sex, race, or age. We assume the possible values for A always include a_0 and a_1, representing a baseline group and a marginalised group, respectively, though the domain of A may contain more values. For example, if A represents racial category, A may take $k \geq 2$ values, with a_0, a_1 representing individuals categorised as white and black respectively.

Following the influence diagram literature (Howard and Matheson 1984), we represent $\hat{Y}$ with a square node and U with a hexagonal node, since $\hat{Y}$ can be viewed as a decision optimizing the function f_U. We also adopt the term *utility variable* to refer to U.

An example of SL graph representing a hiring test prediction setting is given in Figure 1. The training data consists of one input feature D, which represents the candidates degree, and a label Y, which represents whether the candidate passes or fails. The graph also include a variable that is not accessible to the predictor, namely a sensitive attribute A, which represents gender. This reflects a scenario in which the sensitive attribute is not available to the developer, or the developer has chosen not to include it as in input, for example due to legal reasons. In this example, all inputs to Y are also inputs to $\hat{Y}$, but in general this may not be the case (see later examples).

For an SL SCM $\mathcal{M}$, we can consider different predictors $\pi : \mathrm{dom}(\mathbf{Pa}^{\hat{Y}}) \to \mathrm{dom}(\hat{Y})$ (where dom denotes the possible outcomes of a set of variables) by replacing the structural function $f_{\hat{Y}}$ with π, which results in a modified SCM $\mathcal{M}_\pi$. A predictor π is *optimal* if it maximizes the expected value of the utility variable $\mathbb{E}(f_U(Y, \hat{Y}))$.

3 Defining Introduced Unfairness

We propose quantifying introduced unfairness with the following approach: (i) select an appropriate measure of unfairness applicable to both $\hat{Y}$ and Y, and (ii) calculate the difference in unfairness between $\hat{Y}$ and Y. A natural choice of unfairness measure is *total variation*, a generalisation of demographic disparity, which measures the difference in average outcome between different values of the sensitive attribute.

Definition 2 (Average total variation; Zhang and Bareinboim 2018b)**.** *The* average total variation *(ATV) on a real-valued variable V is the difference in the expected value of V between the baseline and marginalised group:*

$$ATV(V) = \mathbb{E}(V \mid A = a_1) - \mathbb{E}(V \mid A = a_0).$$

We define the new concept *introduced total variation* as the difference in magnitude of ATV between $\hat{Y}$ and Y.

Definition 3 (Introduced total variation)**.** *In an SL SCM with real-valued Y and $\hat{Y}$, the* introduced total variation *(ITV) is:*

$$ITV = |ATV(\hat{Y})| - |ATV(Y)|.$$

When ITV is positive/zero/negative we will respectively say that there is introduced, reproduced*, or* reduced *total variation.*

We illustrate ITV on a hiring test prediction example represented by the SL graph of Figure 1.

Example: Hiring test prediction. A model predicts job applicants' outcomes on a hiring test using their degree D — either 'maths' or 'statistics'. Degree is in turn affected by the sensitive attribute gender (A). The loss U depends on the true label $Y \in \{0, 1\}$ (representing fail/pass) and on its prediction $\hat{Y}$.

Suppose that 80% of male applicants have degrees in maths (20% in statistics), while 20% of female applicants have degrees in maths (80% in statistics). Performance on the test is such that $P(Y = 1 \mid D = maths) = 51\%$, $P(Y = 1 \mid D = stats) = 49\%$ (otherwise $Y = 0$). This gives $ATV(Y) = 0.012$.

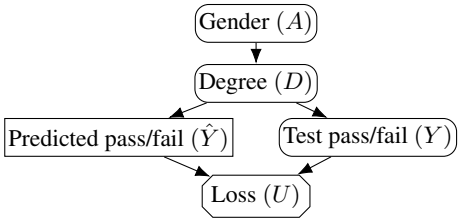

Figure 1: SL graph representing a hiring test prediction setting.

Version 1: For an example with $ITV = 0$, suppose mean squared error were used, and $\hat{Y}$ is a value in $[0, 1]$ representing the probability of passing the test. The optimal predictor would be $f_{\hat{Y}}(maths) = 0.51$, $f_{\hat{Y}}(stats) = 0.49$, yielding $ATV(\hat{Y}) = 0.012$ and therefore $ITV = 0$.

Version 2: Suppose instead that zero-one loss is used. Then the optimal predictor is $f_{\hat{Y}}(maths) = 1, f_{\hat{Y}}(stats) = 0$, yielding $ATV(\hat{Y}) = 0.6$, and therefore a high introduced total variation $ITV = 0.588$, due to the fact that, while female applicants perform only slightly worse than male applicants with respect to Y, their predictions $\hat{Y}$ are vastly lower.

The existence of models with $ITV > 0$ (including in cases where $ATV(Y) = 0$, as in the later music test example) offers a new perspective on the adage that *unfair labels lead to unfair models*: unfair labels may lead to an unfair model, but sometimes, unfairness may originate exclusively, or predominantly from other parts of the training process.

4 ITV, Separation, and Sufficiency

Another way to think about introduced disparities is to ask about reproduced and introduced *dependencies*. These are captured by the existing notions of sufficiency and separation. Indeed ITV is related to the absence of *separation*, a generalization of *equalized odds* to non-binary variables.

Definition 4 (Separation; Barocas, Hardt, and Narayanan 2019)**.** *The random variables $(\hat{Y}, A, Y)$ satisfy* separation *if $\hat{Y}$ is independent of A conditioned on Y, i.e. $\hat{Y} \perp A \mid Y$.*

Absence of separation means that the model has added a *dependence* between A and $\hat{Y}$, that was not present between A and Y. In contrast, ITV asks whether the model has added to the *disparity* in $\hat{Y}$ (as measured by total variation), compared to that in Y. Thus, they both detect whether some effect has been introduced by the model. While lack of separation indicates that the model has introduced some new dependence, ITV captures whether this manifests as an increase (or decrease) in variation between groups. That is, ITV measures the group level impact resulting from the introduction of some new dependency by the model.

For example, in the test prediction examples (Figure 1), separation is not satisfied in either version. But in version 2 (where all statisticians are rejected), we see a large introduced variation. In version 1 (where statisticians are given slightly lower scores), we have $ITV = 0$.

In the binary case, separation prevents introduced variation, as established next by Proposition 5. The converse is not true: it is possible for a model to lack separation while there is a reduced or reproduced variation. For example, version 1 of Figure 1 lacks separation and has $ITV = 0$.

Proposition 5. *Let $\mathrm{dom}(Y) = \{0,1\}, \mathrm{dom}(\hat{Y}) \subseteq [0,1]$, and $\mathrm{dom}(A) \supseteq \{a_0, a_1\}$, where a_0, a_1 are the baseline and marginalised groups. Then separation implies $ITV \leq 0$, i.e. there is not introduced total variation.*

Proof. $|ATV(\hat{Y})|$

$$
\begin{aligned}
&= |\mathbb{E}(\hat{Y}|a_1) - \mathbb{E}(\hat{Y}|a_0)| \\
&= |\textstyle\sum_y P(y|a_1)\mathbb{E}[\hat{Y}|a_1, y] - \sum_y P(y|a_0)\mathbb{E}[\hat{Y}|a_0, y]| \\
&= |\textstyle\sum_y P(y|a_1)\mathbb{E}[\hat{Y}|a_1, y] - \sum_y P(y|a_0)\mathbb{E}[\hat{Y}|a_1, y]| \\
&\qquad\qquad \text{(by separation)} \\
&= |\big(P(Y=1|a_1) - P(Y=1|a_0)\big)\mathbb{E}[\hat{Y}|a_1, Y=1] \\
&\quad - \big(P(Y=1|a_1) - P(Y=1|a_0)\big)\mathbb{E}[\hat{Y}|a_1, Y=0]| \\
&\qquad\qquad (Y \text{ is binary}) \\
&= |\big(P(Y=1|a_1) - P(Y=1|a_0)\big) \\
&\quad .\big(\mathbb{E}[\hat{Y}|a_1, Y=1] - \mathbb{E}[\hat{Y}|a_1, Y=0]\big)| \qquad \text{(factor)} \\
&\leq |P(Y=1 \mid a_1) - P(Y=1 \mid a_0)| \qquad (\text{as } 0 \leq \hat{Y} \leq 1) \\
&= |ATV(Y)|. \qquad \square
\end{aligned}
$$

We also establish the relationship between *sufficiency* and ITV. Sufficiency generalises the notion of *predictive parity*, and is closely related to the notion of *calibration by group* (Barocas, Hardt, and Narayanan 2019). Sufficiency means that the predictor $\hat{Y}$ fully captures the dependencies between A and Y (but does not prohibit additional dependencies being introduced by the model).

Definition 6 (Sufficiency; Barocas, Hardt, and Narayanan 2019)**.** *The random variables $(\hat{Y}, A, Y)$ satisfy* sufficiency *if Y is independent of A conditioned on $\hat{Y}$, i.e. $Y \perp A \mid \hat{Y}$.*

If sufficiency holds, a model may still introduce additional variation. In fact, sufficiency prevents *reduced* variation in the binary case:

Proposition 7. *Let $\mathrm{dom}(\hat{Y}) = \{0,1\}, \mathrm{dom}(Y) \subseteq [0,1]$, and $\mathrm{dom}(A) \supseteq \{a_0, a_1\}$, where a_0, a_1 are the baseline and marginalised groups. Then sufficiency implies $ITV \geq 0$, i.e. there is not reduced total variation.*

Proof. Swap Y and $\hat{Y}$ in the proof of Proposition 5. $\square$

In Appendix A, we consider a related measure, *introduced mutual information*, which can be applied to cases where $\hat{Y}$ and Y are categorical or continuous. Analogous results to Propositions 5 and 7 hold in this more general setting. As a corollary, we prove that it is often impossible to satisfy sufficiency and independence requirements simultaneously, even when more granular definitions are used (e.g. that capture the degree of separation).

5 Incentives for ITV

Under what circumstances will introduced variation arise? Since an arbitrary predictor can introduce variation in almost any setting, we focus on predictors that have been trained to optimality in their given setup. In other words, we ask when introduced variation is *incentivised*.

To specify the graphical criteria, we use the well-known concept of d-separation, which identifies conditional independencies based on the paths between variables.

Definition 8 (d-separation; Verma and Pearl 1988)**.** *A* path *V_1 --- V_k is a sequence of distinct nodes $V_1, ..., V_k, k \geq 0$ such that every pair of consecutive nodes is connected by an edge $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$. When three consecutive nodes in a path have converging edges $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$, we call V_i a* collider*. A path p is said to be* blocked *by the conditioning set $\boldsymbol{Z} \subseteq \boldsymbol{V}$ if p contains a non-collider W in $\boldsymbol{Z}$ or a collider W that is neither equal to, nor an ancestor of, any $Z \in \boldsymbol{Z}$. For disjoint sets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$, the conditioning set $\boldsymbol{Z}$ is said to* d-separate *$\boldsymbol{X}$ from $\boldsymbol{Y}$, if and only if $\boldsymbol{Z}$ blocks every path from a node in $\boldsymbol{X}$ to a node in $\boldsymbol{Y}$. Sets that are not d-separated are called* d-connected*.*

If X and Y are d-separated by Z, then X and Y are conditionally independent given Z in any SCM compatible with the graph, i.e. $P(X \mid Y, Z) = P(X \mid Z)$ (Verma and Pearl 1988). A consequence of d-separation of particular value to us, is that it can be used to establish which features can be useful to an optimal predictor.

Definition 9 (Requisite feature; Lauritzen and Nilsson 2001)**.** *In an SL graph, a feature $W \in \boldsymbol{Pa}^{\hat{Y}}$ is* requisite *if it is d-connected to U conditional on $\boldsymbol{Pa}^{\hat{Y}} \cup \{\hat{Y}\} \setminus \{W\}$. Let $req(\boldsymbol{Pa}^{\hat{Y}})$ denote the set of requisite features.*

Lemma 10 (Fagiuoli and Zaffalon 1998; Shachter 2016)**.** *Every SL SCM has an optimal predictor π that only depends on requisite features, i.e. $\mathbb{E}(\hat{Y} \mid \boldsymbol{Pa}^{\hat{Y}}) = \mathbb{E}(\hat{Y} \mid req(\boldsymbol{Pa}^{\hat{Y}}))$.*

5.1 Arbitrary loss functions

We begin with a graphical criterion for when ITV may incentivised under arbitrary loss functions.

Theorem 11 (Introduced total variation criterion)**.** *An SL graph $\mathcal{G}$ is compatible with an SCM $\mathcal{M}$ in which all optimal predictors have $ITV > 0$ iff there is a requisite feature $W \in \boldsymbol{Pa}^{\hat{Y}}$ that is d-connected to A.*

Proof. We first show that the criterion is sound. Suppose that A is not d-connected to any requisite feature. By Lemma 10, there exists an optimal predictor π that only responds to requisite features. Then

$$\begin{aligned}
\mathbb{E}_\pi(\hat{Y} \mid a) &= \mathbb{E}(\mathbb{E}(\hat{Y} \mid \mathbf{Pa}^{\hat{Y}}, a) \mid a) && \text{(total expectation)} \\
&= \mathbb{E}(\mathbb{E}(\hat{Y} \mid \mathbf{Pa}^{\hat{Y}}) \mid a) && \text{(since } \hat{Y} \perp A \mid \mathbf{Pa}^{\hat{Y}}\text{)} \\
&= \mathbb{E}(\mathbb{E}(\hat{Y} \mid req(\mathbf{Pa}^{\hat{Y}})) \mid a) && \text{(by Lemma 10)} \\
&= \mathbb{E}(\mathbb{E}(\hat{Y} \mid req(\mathbf{Pa}^{\hat{Y}}))) && \text{(by assumption)} \\
&= \mathbb{E}(\hat{Y}). && \text{(total expectation)}
\end{aligned}$$

Therefore $ATV(\hat{Y}) = 0$. It follows that $ITV \leq 0$, i.e. there is no introduced total variation.

For the completeness direction of the proof, in Appendix B we construct a compatible SCM with $ITV > 0$ for any SL graph in which A is d-connected to a requisite feature. □

For example, in the SL graph of Figure 1, $D \in \mathbf{Pa}^{\hat{Y}}$ is a child of A and is d-connected to U conditioned on $\hat{Y}$, and thus the graph satisfies the ITV criterion. This graph is therefore compatible with an ITV incentive, as verified for the particular model stated in version 2, where $ITV = 0.588 > 0$ for the only optimal predictor.

The ITV criterion can be broken down into two conditions, each with an easily interpreted meaning. The first condition says that it is only possible for a predictor to introduce total variation if some feature W can statistically depend on the sensitive attribute A. Otherwise, the total variation of $\hat{Y}$ will be zero (even if the labels Y are strongly dependent on A), and so ITV cannot be positive. The second condition says that ITV can only be incentivised if such a feature W is important for optimal predictions. Indeed, if A is only connected to features that are unimportant for predicting Y, then an optimal predictor may avoid any dependency with A. Note that these conditions can be stated purely in terms of conditional independencies rather than d-separation, so the "only if" part of the theorem can be adapted to settings where we know the joint distribution rather than the graph.

While it is easy to see that if either of the above conditions are not satisfied, then an optimal predictor may avoid introducing total variation, the theorem also establishes the converse: that jointly satisfying the conditions is sufficient for an introduced total variation incentive under some model compatible with the graph. This latter completeness direction of the proof is related to the corresponding completeness proof for d-separation (Geiger, Verma, and Pearl 1990). However, our result is not a corollary of the d-separation result. In particular, the completeness results of d-separation rely on being able to freely specify conditional probability distributions for all nodes. This is not possible here since we are concerned with *optimal* predictors, and so the distribution at $\hat{Y}$ cannot be independently selected (Everitt et al. 2021).

The fact that the conditions of the ITV criterion theorem are easily satisfied indicates that introduced unfairness is possible in a wide range of scenarios.

5.2 P-admissible loss functions

Ideally, we would not just quantify disparities introduced by a system, but would understand what components of the system may be controlled to reduce them. One such component is the training loss function.

As a simple example, if zero-one loss is used, this can lead to a large ITV, because small group differences can be amplified into large differences in "all or nothing" predictions (recall version 2 of the hiring test prediction example). Can this amplification be prevented by choosing a "better behaved" loss function? We investigate an existing class of loss functions that incentivise the predictor to output the expected value of Y given the system inputs.

Definition 12 (P-admissible; Miller, Goodman, and Smyth 1993)**.** *For an SL SCM $\mathcal{M}$ with utility variable U, we say that f_U is a* P-admissible loss function *if $\pi(\boldsymbol{Pa}^{\hat{Y}}) := \mathbb{E}(Y \mid \boldsymbol{Pa}^{\hat{Y}})$ is an optimal predictor.*

Examples of P-admissible loss functions include mean squared error and cross-entropy loss (Miller, Goodman, and Smyth 1993). We now show that for some graphs, using a P-admissible loss function rules out the possibility of an ITV incentive.

Theorem 13 (P-admissible ITV criterion)**.** *An SL graph $\mathcal{G}$ is compatible with an SCM $\mathcal{M}$ for which f_U is P-admissible and all optimal predictors have $ITV > 0$ if in addition to the conditions of Theorem 11, $A \notin \boldsymbol{Pa}^{\hat{Y}}$ and A is d-connected to U conditioned on $\boldsymbol{Pa}^{\hat{Y}}$.*

Proof. Beginning with the soundness direction, consider the case in which one of the conditions does not hold. If the Theorem 11 condition does not hold, then $\mathcal{G}$ is not compatible with an SCM $\mathcal{M}$ with an ITV incentive under any loss function, including P-admissible ones.

If the extra condition of Theorem 13 does not hold, then for any $a \in \mathrm{dom}(A)$:

$$\begin{aligned}\mathbb{E}(\hat{Y} \mid a) &= \mathbb{E}(\mathbb{E}(Y \mid \mathbf{Pa}^{\hat{Y}}) \mid a) && \text{(by P-admissibility)}\\ &= \mathbb{E}(\mathbb{E}(Y \mid \mathbf{Pa}^{\hat{Y}}, a) \mid a) && \text{(see below)}\\ &= \mathbb{E}(Y \mid a). && \text{(law of total probability)}\end{aligned}$$

The second equality holds if either (a) $A \in \mathbf{Pa}^{\hat{Y}}$, or (b) A is d-separated to Y conditioned on $\mathbf{Pa}^{\hat{Y}}$. Therefore since we have assumed that the extra condition does not hold, the second equality follows. Since $\mathbb{E}(\hat{Y} \mid a) = \mathbb{E}(Y \mid a)$ for all a it follows that $ITV = 0$.

See Appendix B for the completeness direction. Note that the completeness proof relies on A being able to take at least three different values – if A is forced to be binary, an even stricter graphical criterion may hold. □

The conditions of Theorem 13 are more restrictive than that in Theorem 11, as they additionally require A to be a non-parent of $\hat{Y}$ and d-connected to the utility variable. This can be interpreted as follows: positive ITV requires that adding A as a feature would provide some additional *value of information* (Howard 1966; Everitt et al. 2021) given the available features. Theorem 13 implies that for any model whose graph does not satisfy the more restrictive criterion, using a P-admissible loss removes the undesired incentive.

For example, Figure 1 does not satisfy the extra condition of Theorem 13 as the only path from A to U is blocked by D. Indeed, if a non-P-admissible loss function is replaced with mean squared error, the optimal policy becomes $f_{\hat{Y}}(maths) = 0.51, f_{\hat{Y}}(stats) = 0.49$ and $ITV = 0$, as we saw in version 1. We now present a case that satisfies the conditions of Theorem 13, and allows for $ITV > 0$ even under P-admissible loss.

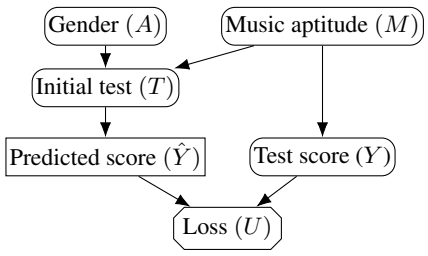

Figure 2: SL graph representing a music test prediction setting.

Example: Music test prediction. Consider the SL graph in Figure 2, which represents a music test scenario in which a model is trained to predict the outcome $Y \in \{0, 1\}$ of a test taken at the end of a music course (adapted from Chiappa and Isaac 2019). The prediction is based only on an initial test outcome $T \in \{0, 1\}$, which has a gender bias. Assume equal numbers of females and males, both with equal numbers of low and high musical aptitude (represented by $M = 0, 1$ respectively), take tests T and Y. Suppose that 95% of individuals with high aptitude ($M = 1$) pass the final test ($Y = 1$), compared to 5% of individuals with low aptitude ($M = 0$). Suppose that 90% of high aptitude females pass the *initial* test ($T = 1$), compared to 100% of males. Low aptitude individuals also pass test T with 5% probability.

This scenario satisfies the conditions of Theorem 13, since A is d-connected to the feature T, which is requisite (it is d-connected to U), and conditioning on T opens the path from $A \notin \mathbf{Pa}^{\hat{Y}}$ to U (via M and Y). As expected, we find that the optimal predictor does have an introduced total variation under both zero-one loss (the optimal predictor is $f_{\hat{Y}}(1) = 1, f_{\hat{Y}}(0) = 0$, and $ITV = 0.05$) and under mean squared error (it can be shown that the optimal predictor is $f_{\hat{Y}}(1) = 0.905, f_{\hat{Y}}(0) = 0.095$, and $ITV = 0.0405$).

This also shows that an introduced variation can arise even when the training labels are completely unbiased, i.e. when $ATV(Y) = 0$ and indeed A is independent of Y. Wang et al. (2019) describe a similar dynamic in the context of image classification.

Removing an ITV incentive. The P-admissible ITV criterion of Theorem 13 generalises previous observations that imposing fairness through unawareness, namely avoiding explicit use of the sensitive attributes to form decisions, might actually introduce unfairness. Specifically, Chiappa and Isaac show that when the variables in Figure 2 are linearly related, a linear least-squares predictor without access to A will result in a biased $\hat{Y}$, whilst access to A will mean the predictor "strips off" the bias in T from $\hat{Y}$. The following corollary of Theorem 13 generalises this result to arbitrary graphs, non-linear relationships among variables, and a more general class of predictors.

Corollary 14. *If the sensitive variable A is available as a feature to the predictor, then $\mathcal{G}$ is not compatible with an ITV incentive under P-admissible loss functions.*

Corollary 14 offers a general method for preventing $ITV > 0$: add A as a feature and use a P-admissible loss function. For example, adding A as a feature to Figure 2 (i.e. adding a link from A to $\hat{Y}$), breaks the extra condition of Theorem 13. Indeed, for the example given, it can be shown that if we add A as feature, the optimal policy results in $ITV = 0$.

As with any technique to ensure fairness, making A available as a feature should not be done without an understanding of the context. In particular, since $ITV = 0$ is a specific group-level measure, it does not come with individual-level guarantees. In the music test example, as the initial test has lower accuracy for women, women who pass the initial test receive a slightly lower prediction when A is used explicitly compared to when it is not. Under mean squared error, when A is a added as a feature the optimal predictor yields $f_{\hat{Y}}(T = 1, A = \text{woman}) = \mathbb{E}(Y \mid T = 1, A = \text{woman}) = 0.903$. When A is not a feature, $f_{\hat{Y}}(T = 1) = \mathbb{E}(Y \mid T = 1) = 0.905$. This negative effect is offset by the higher score given to women who failed the test: if A is a feature, the optimal predictor yields $f_{\hat{Y}}(T = 0, A = \text{woman}) = \mathbb{E}(Y \mid T = 0, A = \text{woman}) = 0.14$. If A is not a feature, $f_{\hat{Y}}(T = 0) = \mathbb{E}(Y \mid T = 0) = 0.095$. This may be perceived as unfair by the high aptitude women who passed the test T.

We note that in the case of classification, requiring a discrete deterministic result will mean that a P-admissible loss function cannot be used. For instance, if mean squared error is used to produce $\hat{Y} = p \in [0, 1]$, but a binary accept/reject is required, then thresholding (e.g. at 0.5) reduces to the zero-one loss case, and may give $ITV > 0$, even if the Theorem 13 criteria are met. In this case, randomising the result (accepting with probability p) preserves the result. However, our results do not rely on randomness in general. E.g. consider situations where the prediction task is inherently continuous, such as a sum of money paid out to an insurance customer. Then there would be no need to randomise (or threshold) and the results would be preserved.

6 Path-specific Introduced Effects

In previous sections, we have examined introduced unfairness for statistical definitions of unfairness. However, causal definitions can offer a more fine-grained understanding of unfairness (Pearl 2009; Kusner et al. 2017; Zhang, Wu, and Wu 2017; Loftus et al. 2018; Chiappa 2019; Chiappa et al. 2020; Oneto and Chiappa 2020). In this section, we consider a notion of introduced unfairness based on causal effects restricted to certain paths, referred to as *path-specific effects* (PSEs). We first recap causal interventions and path-specific effects, and then adapt the idea of introduced unfairness to define *introduced path-specific effects*, and illustrate how they may be used to examine the source of an introduced effect.

6.1 Background on path-specific effects

As well as allowing us to investigate the result of conditioning on a particular variable, SCMs also allow us to investigate the result of *intervening* on a particular variable, to answer causal questions. Formally, an intervention in an SCM $\mathcal{M}$ consists in setting a variable X to the value x by replacing the structural function f_X with a constant function $f_X = x$. The variables in the modified model are referred to as V_x. Interventions on X only alter the values of variables descending from X, so $V_x = V$ for non-descendants of X. *Path-specific* interventions are a more targeted type of intervention, that only propagate along specific paths. While global interventions allow us to reason about the total causal effect of a variable, for example to answer the question, "What effect did being male have on being hired in a job application?", path-specific interventions enable us to reason about the effect along a subset of paths, for example to answer a more fine-grained question, "What effect did indicating male associated hobbies on resumes have on being hired in a job application?".

Definition 15 (Path-specific effect; Pearl 2001)**.** *For a given edge-subgraph $\mathcal{P}$ specifying a set of paths in an SCM $\mathcal{M}$, let $\mathcal{M}_{\mathcal{P}}$ be a modified version of $\mathcal{M}$ in which all function inputs not in $\mathcal{P}$ are kept fixed at a baseline value $A = a_0$. That is, replace each structural function $f_X(V^1, \ldots, V^k)$ in $\mathcal{M}$ with the function $(f_X)'$, equal to f_X, except that if $V^i \to X$ is not in $\mathcal{P}$, then when evaluating at $\boldsymbol{\varepsilon}$ the argument V^i is replaced with the constant $V^i_{a_0}(\boldsymbol{\varepsilon})$. The* path-specific response $V_{\mathcal{P}(a_0 \to a_1)}$ *is defined as V_{a_1} in the model $\mathcal{M}_{\mathcal{P}}$. The* path-specific effect *(PSE) on a real-valued variable V is:*

$$PSE(V) = \mathbb{E}(V_{\mathcal{P}(a_0 \to a_1)}) - \mathbb{E}(V_{a_0}).$$

6.2 Auditing ML system outputs for fairness – a risk when labelling paths to $\hat{Y}$ as fair/unfair

Path-specific effects can be used to inform judgements about whether a decision policy is unfair. For example, in the case of Berkeley's alleged sex bias in graduate admissions, the original analysis considered direct effects (*Gender* $\to$ *Outcome*) to be unfair, but indirect effects via (*Gender* $\to$ *Department* $\to$ *Outcome*) to be fair (Bickel, Hammel, and O'Connell 1975; Pearl 2009). This approach assumes that societal considerations can be used to label paths between the sensitive variable A and the outcome as fair (or "justified") or unfair, and outcomes are declared unfair if (significant) effects are found along any unfair path.

Suppose instead that the aim is to audit the fairness of a trained ML system, by investigating the system outputs. While understanding which paths are responsible for disparate $\hat{Y}$ is crucial, here we show that the training process must also be taken into account before attempting to label paths from a sensitive variable A to $\hat{Y}$ as fair or unfair.

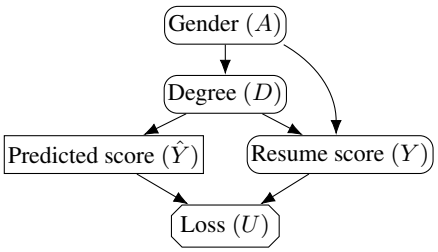

Figure 3: Penalising female degrees example.

Example: Penalising female dominated degrees. In the SL graph of Figure 3, a system $\hat{Y}$ is used to score resumes based on the applicant's degree D. The system is trying to emulate scores Y given by humans, that also directly depend on applicant's gender A. If we deem degree to be a reasonable decision criterion for the job in question, it is tempting to label the path $A \to D \to \hat{Y}$ as fair, and therefore to conclude that the decision given by $\hat{Y}$ must be fair.

Assume for simplicity that there are only two degrees, maths and statistics, and that they are equally valuable, with reviewers giving them both a score of 5. In addition, suppose that male and female applicants are (unfairly) given an additional score of 1 and -1 respectively. Moreover, suppose that 80% of mathematicians are male, while 80% of statisticians are female, and that mean squared error is used as the loss function. Then the optimal predictor is $f_{\hat{Y}}(maths) = \mathbb{E}(Y \mid D = maths) = 5 + 0.6 = 5.6$ and $f_{\hat{Y}}(stats) = \mathbb{E}(Y \mid D = stats) = 5 - 0.6 = 4.4$.

Thus statistics is given a lower score by the model, even though it is just as valuable as mathematics, *because it has a higher percentage of women*. Knowing this, we may instead conclude that these decisions at $\hat{Y}$ are *not* fair, because the path-specific effect of A on $\hat{Y}$ via D is stronger than the corresponding path-specific effect on Y.

To formalise this type of situation, we define path-specific introduced effects.

Definition 16 (Path-specific introduced effect)**.** *Let $\mathcal{M}$ be an SL SCM with real-valued variables Y and $\hat{Y}$. Let $\mathcal{P}$ denote an edge-subgraph with paths from A to Y and A to $\hat{Y}$. Then the* path-specific introduced effect *(PSIE) is defined as*

$$PSIE_{\mathcal{P}} = |PSE_{\mathcal{P}}(\hat{Y})| - |PSE_{\mathcal{P}}(Y)|.$$

In particular, if $\mathcal{P}$ consist of all directed paths of the form $A \dashrightarrow X \dashrightarrow \hat{Y}$ and $A \dashrightarrow X \dashrightarrow Y$, then $PSIE_{\mathcal{P}}$ is the path-specific introduced effect via variable X.

For example, in Figure 3, let $\mathcal{P} := \{A \to D \to \hat{Y}, A \to D \to Y\}$. It can be shown that $PSIE_{\mathcal{P}} = 0.72 - 0 = 0.72$ and conclude that there is a PSIE via D. In other words, D is carrying a (problematic) amplified effect, as a result of the path partially reproducing the spurious effect along the path D $\leftarrow A \to Y$. As the effect along $A \to D \to Y$ may be considered fair in this example, we may say that the effect of A on $\hat{Y}$ via D is greater than that justified by the equivalent path to Y. Note that looking at the introduced total variation would not identify this phenomena, as $ITV = -1.28 < 0$ in this example.

In this example, the training labels Y are unfair. However, a PSIE via a variable X can also occur in cases where Y is considered fair. For example, assume that, in Figure 3, $A \to Y$ is replaced with $A \to$ *Coding Experience* $\to Y$ and that coding is deemed fair (despite more men having coding experience) since it is relevant for the job. Then D can nonetheless still carry a problematic PSIE exactly as before – statistics will be still be penalised for having a higher percentage of women.

These examples highlight the importance of taking PSIE into account when decided whether an effect along a path is fair or unfair.

Relationship to proxy unfairness. Kilbertus et al. (2017) define proxy discrimination as arising if a causal path from A to a decision is blocked by a variable deemed to be a proxy (e.g. someone's name may be considered a proxy for their gender), but does not describe how to ascertain whether a variable should be considered a proxy. PSIE gives information that can help judge whether a variable is a (problematic) proxy, namely whether it carries an amplified effect from A. In the examples described above, the seemingly harmless degree carries an amplified effect; it thus acts as a harmful proxy for gender.

6.3 Enforcing ML system outputs to be fair - A risk when reducing unfair PSEs.

Several approaches in the literature are based on enforcing path-specific effects or counterfactual extensions that are considered problematic in the data not to be transferred to the system (e.g. Nabi and Shpitser (2018); Chiappa (2019)). These approaches implicitly assume that the prediction model and training data share the same underlying causal structure, and ensure that the effect on any path corresponding to an unfair path underlying the data is reduced, either by constraining the objective during training (Nabi and Shpitser 2018) or by performing a path-specific counterfactual prediction at test time (Chiappa 2019). However, the discussion above indicates that effects on paths that are deemed fair also need to be considered. Specifically, consider Figure 3, but with an additional direct path $A \to \hat{Y}$, so that the causal structure underlying Y and $\hat{Y}$ are the same. Ensuring that the effect along the harmful path $A \to Y$ is not reproduced as $A \to \hat{Y}$ is not sufficient to ensure fairness: the effect via $A \to D \to \hat{Y}$ needs to also be understood. Any method that constrains the learning to reduce the effect along "unfair" paths risks transferring this effect to a "fair" path, such as the one through D. Methods that only learn the causal model underlying the data without such constraints might still carry some risk. PSIE can be used to formalise these risks. In addition, our formalism can be used to understand when such an amplified effect (naturally) arises as a consequence of optimality, i.e. when such an effect is incentivised.

6.4 Incentives for PSIE

As we did for ITV in the previous section, here we ask: when may PSIE be incentivised by a training setup? The conditions are very similar to Theorem 11.

Theorem 17 (PSIE graphical criterion)**.** *An SL graph $\mathcal{G}$ is compatible with an SCM $\mathcal{M}$ in which all optimal predictors have $PSIE_{\mathcal{P}} > 0$ iff there is some path $p \in \mathcal{P}$ of the form $A \dashrightarrow W \to \hat{Y}$ where $W \in req(\boldsymbol{Pa}^{\hat{Y}})$ is a requisite feature.*

Proof. Let $\mathcal{G}$ be an SL graph $\mathcal{G}$ and $\mathcal{P}$ an edge-subgraph that includes no path $A \dashrightarrow W \to \hat{Y}$ via a requisite feature W. By Lemma 10, for any compatible SCM there is an optimal predictor that only depends on requisite features. Under this predictor $PSE_{\mathcal{P}}(\hat{Y}) = 0$. Since $PSIE_{\mathcal{P}} \leq PSE_{\mathcal{P}}(\hat{Y})$, we have established that the graphical criterion is sound. A proof of completeness can be found in Appendix C. □

7 Empirical Results

Our graphical criteria give conditions under which a graph is *compatible* with ITV, or PSIE. But do these arise in practice? For random distributions, d-connectedness almost always implies conditional dependence (Meek 1995). Therefore the requisite feature W required by the ITV and PSIE criteria will almost always have a dependency with A and U when the ITV criteria are satisfied (under random distributions). However, this does not necessarily imply positive ITV. In particular, our completeness results establish the existence of some model where ITV is positive, but not how common positive ITV is. We address this second question empirically.

PyCID is an open source Python library for graphical models of decision-making (Fox et al. 2021). Using PyCID's method for generating random graphs, we sample SL SCMs with 6 nodes and random distributions, that satisfy the graphical criteria of Theorems 11 and 13. Out of 1000 samples for each, we found that 20% of the models satisfying the graphical criteria of Theorem 11 have ITV greater than 0.01 under zero-one loss, while 16% of the models satisfying the graphical criteria of Theorem 13 have ITV greater than 0.01

under P-admissible loss. The results did not appear particularly sensitive to variations in the number of nodes or the edge density of the random graph. The results can easily be reproduced by the linked colab[1], which also shows how the examples discussed above can be analysed using PyCID.

8 Related Work

Statistical approaches. In Section 3, we discuss the relationship between ITV and separation and sufficiency, for example that lack of separation means that a dependency has been introduced (in the binary case), whereas ITV measures whether this manifests as any increased (or decreased) disparity.

Causal approaches. The ability to account for the complex patterns that underlie the data generation process makes causal models a powerful tool for reasoning about fairness. As such, causal models are increasingly used both for measuring and alleviating unfairness in ML systems (Chiappa and Isaac 2019; Creager et al. 2020; Loftus et al. 2018; Nabi, Malinsky, and Shpitser 2019; Plecko and Meinshausen 2020; Qureshi et al. 2016; Russell et al. 2017; Zhang and Bareinboim 2018b; Zhang, Wu, and Wu 2017). The idea of inferring the presence of unfairness in data with path-specific effects and counterfactuals dates back to Pearl (2009) and Pearl, Glymour, and Jewell (2016). Kilbertus et al. (2017); Kusner et al. (2017) and Nabi and Shpitser (2018) develop approaches for training ML systems that achieve a coarse-grained version of path-specific fairness, counterfactual fairness, and path-specific fairness respectively. The following work of Chiappa (2019) and Chiappa et al. (2020) introduces general methods for achieving path-specific counterfactual fairness, while Wu et al. (2019) discuss identification issues, and how to compute path-specific counterfactuals.

Attempts to describe the relation between the data and model outputs Y and $\hat{Y}$ have appeared in some of these works, with the goal of elucidating limitations of statistical fairness definitions at a high level (Chiappa and Isaac 2019; Kilbertus et al. 2017). Zhang and Bareinboim (2018a) is the first work to more thoroughly characterise the causal connection between the two variables, by linking the equalised odds criterion to the underlying causal mechanisms. This work differs from ours in several ways. Our goal in characterising the relation between Y and $\hat{Y}$ is not to connect statistical fairness definitions to the underlying data generation mechanisms, but to formalise the notion that models may introduced or amplify causal effects not present in the training labels. In addition, rather than reasoning about a trained model for $\hat{Y}$, we also incorporate the training mechanism by considering the necessary behaviour of optimal predictors (Everitt et al. 2021). This enables us to characterise when policies are incentivised to introduce or amplify disparities that were not present in the training labels.

[1]https://github.com/causalincentives/pycid/blob/master/notebooks/Why_fair_labels_may_yield_unfair_models_AAAI_22.ipynb

Amplified disparity in context. There is also a broader literature that investigates the relationship between biased labels and biased models for particular applications, such as object recognition (Wang et al. 2019; Zhao et al. 2017). For example, this may result from the fact that even if Y and A are independent, some features X might be correlated with both A and Y, inducing a correlation between A and $\hat{Y}$ (Wang et al. 2019). This can be seen as an example of ITV. In contrast to these works, we seek a theoretical understanding of when introduced disparity will arise, particularly in decision-making settings about individuals.

9 Discussion

Applicability of incentive criteria. The graphical criteria can be used to analyse the potential incentives of a system that is yet to be built, or for which we lack access to model outputs for other reasons (e.g. a proprietary system). The necessary graphical knowledge may come from domain expertise, previous studies, or data. For example, a developer or auditor may know that *Ability* is a joint ancestor of *Test score* and *Job performance*, even if they are unable to measure this directly. Using only this qualitative, "graphical" knowledge, our results establish how potential incentives for ITV and PSIE can be assessed. A weakness is that incentives can only be excluded, not confirmed. For the latter task, access to the data distribution is needed.

Measuring introduced unfairness. When we have access to the model's outputs, we may wish to measure its introduced unfairness. This is usually possible for ITV (given appropriate data), as it is defined in terms of conditional probabilities, which can be easily estimated if the variables are observed. Measuring PSIE is often more challenging, as it is defined in terms of PSEs, which typically require knowledge of the causal graph, and sometimes even the exact structural functions, to calculate. The exact conditions for identifying the PSEs are given by Theorems 4 and 5 of Avin, Shpitser, and Pearl (2005) for Markovian models (i.e. models in which every exogenous variable is independent and in the domain of at most one function f_V), and in Theorems 3 and 4 of Shpitser (2013) for non-Markovian models. Alternative definitions for PSE can also be used in the PSIE definition. See (Shpitser 2013) for details of a more readily estimated (though less general) variant.

Limitations and risks. In addition to the limitations of graphical models (e.g. the sensitivity of results to assumptions), our graphical criteria results pertain to *optimal* policies. Trained models may be substantially suboptimal, if the model class is insufficiently powerful, or insufficient training data is used (Miller, Goodman, and Smyth 1993). In addition, our graphical criteria give conditions for *compatibility* with some incentive – meeting the criteria does not guarantee an incentive for all parameterisations. That said, our empirical results show that an incentive does arise a large proportion of the time. Similarly, failing to meet the criteria guarantees that optimal policies without the property exist, but does not guarantee this for all optimal policies.

It is especially important to take account of the limitations of fairness measures because if inappropriately ap-

plied, they could cause a failure to recognise and address actual injustices. We highlight four limitations, starting with the most general: 1) Fairness definitions require us to define and formalise group membership, an exercise that is fraught with practical and ethical difficulties (West, Whittaker, and Crawford 2019; Kohler-Hausmann 2018; Hanna et al. 2020). 2) Definitions of unfairness are liable to miss manifestations of injustice, and aspects of what we mean by unfairness (Kohler-Hausmann 2018). 3) Group fairness definitions may overlook (un)fairness to individuals. (Dwork et al. 2012; Kleinberg, Mullainathan, and Raghavan 2017). 4) Translating a causal effect into a normative fairness judgement is often complex. While we aim to assist with this as in the discussion around Figure 3, we do not claim to completely resolve this problem.

Findings. In this paper we have proposed new definitions to formalise the introduction of variation (ITV) and causal effects (PSIE) by supervised learning models, and established their graphical criteria. In particular, our analysis revealed that:

- **Incentives depend on the loss function.** In some scenarios in which an introduced effect is incentivised, replacing the loss function with a P-admissible loss function removes the incentive.
- **Correctly labelling paths to $\hat{Y}$ as fair/unfair requires awareness of PSIE.** When considering whether an effect from A to $\hat{Y}$ is unfair, one may need to consider whether variables on the path could carry an unwanted amplified effect.
- **It is difficult to rule out introduced disparity/effects.** The graphical criteria for ITV and PSIE are easily met.
- **Fair training labels do not always yield a fair model.**

Acknowledgements

We would like to thank for their comments, help, and discussions, Ben Coppin, James Fox, Lewis Hammond, William Isaac, Zac Kenton, Claudia Shi, and Chris van Merwijk. This work was supported in-part by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC2015-067.

References

Avin, C.; Shpitser, I.; and Pearl, J. 2005. Identifiability of Path-specific Effects. In *International Joint Conference on Artificial Intelligence*, 357–363.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Bickel, P. J.; Hammel, E. A.; and O'Connell, J. W. 1975. Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, 187(4175): 398–404.

Chiappa, S. 2019. Path-Specific Counterfactual Fairness. In *AAAI Conference on Artificial Intelligence*, 7801–7808.

Chiappa, S.; and Isaac, W. S. 2019. A Causal Bayesian Networks Viewpoint on Fairness. In *E. Kosta, J. Pierson, D. Slamanig, S. Fischer-Hübner, S. Krenn (eds) Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data. Privacy and Identity 2018. IFIP Advances in Information and Communication Technology*, volume 547. Springer, Cham.

Chiappa, S.; Jiang, R.; Stepleton, T.; Pacchiano, A.; Jiang, H.; and Aslanides, J. 2020. A General Approach to Fairness with Optimal Transport. In *AAAI Conference on Artificial Intelligence*, 3633–3640.

Creager, E.; Madras, D.; Pitassi, T.; and Zemel, R. 2020. Causal Modeling for Fairness in Dynamical Systems. In *International Conference on Machine Learning*, 2185–2195.

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

Everitt, T.; Carey, R.; Langlois, E. D.; Ortega, P. A.; and Legg, S. 2021. Agent Incentives: A Causal Perspective. In *AAAI Conference on Artificial Intelligence*, 11487–11495.

Fagiuoli, E.; and Zaffalon, M. 1998. A Note about Redundancy in Influence Diagrams. *International Journal of Approximate Reasoning*, 19: 351–365.

Fox, J.; Everitt, T.; Carey, R.; Langlois, E.; Abate, A.; and Wooldridge, M. 2021. PyCID: A Python Library for Causal Influence Diagrams. In *SciPy*.

Geiger, D.; Verma, T.; and Pearl, J. 1990. d-separation: From Theorems to Algorithms. *Machine Intelligence and Pattern Recognition*, 10: 139–148.

Hanna, A.; Denton, E.; Smart, A.; and Smith-Loud, J. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Conference on Fairness, Accountability, and Transparency*, 501–512.

Hertweck, C.; and Räz, T. 2022. Gradual (In) Compatibility of Fairness Criteria. In *AAAI Conference on Artificial Intelligence*.

Howard, R. A. 1966. Information Value Theory. *IEEE Trans. Systems Science and Cybernetics*, 2(1): 22–26.

Howard, R. A.; and Matheson, J. E. 1984. Influence Diagrams. *The Principles and Applications of Decision Analysis*, 2: 719–763.

Ihara, S. 1993. *Information Theory for Continuous Systems*, volume 2. World Scientific.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems 30*, 656–666.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Innovations in Theoretical Computer Science Conference*.

Kohler-Hausmann, I. 2018. Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination. *Nw. UL Rev.*, 113: 1163.

Koller, D.; and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30*, 4069–4079.

Lauritzen, S. L.; and Nilsson, D. 2001. Representing and Solving Decision Problems with Limited Information. *Management Science*.

Loftus, J. R.; Russell, C.; Kusner, M. J.; and Silva, R. 2018. Causal Reasoning for Algorithmic Fairness. *arXiv:1805.05859*.

Meek, C. 1995. Strong-completeness and Faithfulness in Belief Networks. In *Conference on Uncertainty in Artificial Intelligence*, 411–418.

Miller, J. W.; Goodman, R.; and Smyth, P. 1993. On Loss Functions which Minimize to Conditional Expected Values and Posterior Probabilities. *IEEE Transactions on Information Theory*, 39(4): 1404–1408.

Nabi, R.; Malinsky, D.; and Shpitser, I. 2019. Learning Optimal Fair Policies. In *International Conference on Machine Learning*, 4674–4682.

Nabi, R.; and Shpitser, I. 2018. Fair Inference On Outcomes. In *AAAI Conference on Artificial Intelligence*, 1931–1940.

Oneto, L.; and Chiappa, S. 2020. Fairness in Machine Learning. In Oneto, L.; Navarin, N.; Sperduti, A.; and Anguita, D., eds., *Recent Trends in Learning From Data. Studies in Computational Intelligence*, volume 896, 155–196. Springer, Cham.

Pearl, J. 1986. Fusion, Propagation, and Structuring in Belief Networks. *Artificial intelligence*, 29(3): 241–288.

Pearl, J. 2001. Direct and Indirect Effects. In *Conference in Uncertainty in Artificial Intelligence*, 411–420.

Pearl, J. 2009. *Causality*. Cambridge University Press, 2nd edition.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Plecko, D.; and Meinshausen, N. 2020. Fair Data Adaptation with Quantile Preservation. *Journal of Machine Learning Research*, 21(242): 1–44.

Qureshi, B.; Kamiran, F.; Karim, A.; and Ruggieri, S. 2016. Causal Discrimination Discovery Through Propensity Score Analysis. arXiv:1608.03735.

Russell, C.; Kusner, M.; Loftus, C.; and Silva, R. 2017. When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. In *Advances in Neural Information Processing Systems 30*, 6414–6423.

Shachter, R. D. 2016. Decisions and Dependence in Influence Diagrams. In *International Conference on Probabilistic Graphical Models*, 462–473.

Shpitser, I. 2013. Counterfactual Graphical Models for Longitudinal Mediation Analysis with Unobserved Confounding. *Cognitive Science*, 37(6): 1011–1035.

Verma, T.; and Pearl, J. 1988. Causal Networks: Semantics and Expressiveness. In *Conference on Uncertainty in Artificial Intelligence*, 69–78.

Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.-W.; and Ordonez, V. 2019. Balanced Datasets are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *IEEE/CVF International Conference on Computer Vision*, 5310–5319.

West, S. M.; Whittaker, M.; and Crawford, K. 2019. Discriminating Systems. *AI Now*.

Wu, Y.; Zhang, L.; Wu, X.; and Tong, H. 2019. PC-Fairness: A Unified Framework for Measuring Causality-based Fairness. In *Advances in Neural Information Processing Systems 32*.

Zhang, J.; and Bareinboim, E. 2018a. Equality of Opportunity in Classification: A Causal Approach. In *Advances in Neural Information Processing Systems 32*, 3675–3685.

Zhang, J.; and Bareinboim, E. 2018b. Fairness in Decision-Making – The Causal Explanation Formula. In *AAAI Conference on Artificial Intelligence*, 2037–2045.

Zhang, L.; Wu, Y.; and Wu, X. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *International Joint Conference on Artificial Intelligence*, 3929–3935.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Conference on Empirical Methods in Natural Language Processing*.