Automatic Causal Fairness Analysis with LLM-Generated Reporting

Alessia Berarducci, Eric Rossetto, Alessandro Antonucci, Marco Zaffalon

IDSIA-SUPSI, Lugano, Switzerland
name.surname@supsi.ch

Abstract. AutoML, intended as the process of automating the application of machine learning to real-world problems, is a key step for AI popularisation. Most AutoML frameworks are not accounting for the potential lack of fairness in the training data and in the corresponding predictions. We introduce FAIRMIND, a software prototype aiming to automatise fairness analysis at the dataset level. We achieve that by resorting to the assumptions of the *standard fairness model*, recently proposed by Plečko and Bareinboim. This allows for a sound fairness evaluation in terms of causal effects, based on *counterfactual* queries involving the target, possibly confounders and mediators, and the different values of an input feature we regard as *protected*. Given a dataset and the roles of its variables, the tool implements a closed-form computation of the effects. An LLM is then used, as an interpretative layer—performing no computation—to turn the computed effects into a natural-language report. We achieve that in a zero-shot setup and show by examples the expected advantages with respect to a direct analysis performed by the LLM. To favour applications, extensions to ordinal protected variable and continuous targets and novel decomposition results are also discussed.

Keywords: Causality · Fairness Evaluation · Mediation Analysis · Standard Fairness Model · Natural Language · AutoML.

1 Introduction

In recent years, significant efforts have been devoted to *automated* machine learning (AutoML), where ML algorithms are trained from data with zero or minimal human preprocessing and directly used to perform predictions [1]. The evaluation and mitigation of *fairness* issues also emerged as an important aspect to be considered in the design of these predictive tools [15]. However, to the best of our knowledge, little attention has been paid to fairness within the AutoML community. This situation appears unfortunate, since frameworks such as the causal one proposed by Plečko and Bareinboim [12] already provide a powerful and systematic approach to fairness quantification.

A major goal of this paper is therefore to promote such a counterfactual fairness framework in AutoML settings by reviewing the semantics of the key inferences, together with the corresponding identification formulae. These quantities can be obtained from observational data and their computation naturally embedded within the tool we present. Notably, we also extend the framework to a number of more general cases commonly encountered in applications, including ordinal protected variables and continuous target variables. The resulting quantitative analysis is paired with a *large language*

model (LLM), that provides automated textual reporting based on a zero-shot paradigm guided by a *chain-of-thoughts* scheme. This is in line with the most recent trends in AutoML, where LLM agents provide higher flexibility and usability [7].

Within the Pearlian causal inference framework [11], adapted to algorithmic fairness in [12], multiple works have explored fairness from a theoretical perspective. However, these approaches typically rely on manual modelling decisions and expert interpretation, which somehow contrasts with the AutoML philosophy [15]. Instead of proposing new fairness definitions, we integrate existing causal fairness methods into an automated pipeline. We develop a tool that automatically selects, estimates and explains causal fairness effects with minimal user interventions. Result interpretation is supported by an LLM, enabling non-experts to conduct and understand causal fairness analyses within an AutoML workflow. The work in [8] also describes an implementation of counterfactual fairness metrics, but it only considers aggregated effects and it is not publicly available. Our framework focuses instead on automating effect selection, interpretation and reporting through LLM-assisted pipelines. A web interface to use the framework is available.[1] The code and demonstrative examples are also available.[2]

The paper is organised as follows. Sect. 2 introduces the basic concepts and notation. The standard fairness model is presented in Sect. 3, together with the necessary identification formulae. Sect. 4 presents our extensions of the framework. The automatic reporting based on LLMs is discussed in Sect. 5 together with the FAIRMIND application we release. Conclusions and outlook for future work are in Sect. 6. Proofs are collected in App. A, whilst the prompt used for generating the natural-language report can be found in App. B.

2 Basics

We use uppercase letters for variables, lowercase is used instead for states, and calligraphic for sets of the states. Accordingly, $v \in \mathcal{V}$ is a generic state of V.

Structural Causal Models (SCMs). Following [10], we define an SCM $\mathcal{M}$ (SCMs) as a tuple $\langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P(\boldsymbol{U}) \rangle$ where: (i) $\boldsymbol{V}$ is a set of *endogenous* (observed) variables; (ii) $\boldsymbol{U}$ is a set of *exogenous* (latent) variables distributed according to $P(\boldsymbol{U})$; (iii) $\mathcal{F}$ is a collection of functions, denoted as *structural equations*. Each $V_i \in \boldsymbol{V}$ is determined by a function $f_{V_i} \in \mathcal{F}$, defined as a mapping from a set of exogenous variables $\boldsymbol{U}_i \subseteq \boldsymbol{U}$ and a set of endogenous variables $\mathrm{Pa}_{V_i} \subseteq \boldsymbol{V} \setminus \{V_i\}$ to the domain of V_i. The structural equations induce a *causal graph* $\mathcal{G}$, where an edge (X, Y) exists if and only if $X \in \mathrm{Pa}_Y$. Throughout the manuscript, we focus our attention on *recursive* SCMs, whose causal graph is acyclic. Bi-directed edges ($V_i \dashleftarrow\dashrightarrow V_j$) indicate a common (latent) *confounder* $U \in \boldsymbol{U}$ appearing both in f_{V_i} and f_{V_j}. In the absence of bi-directed edges, i.e., if there is a one-to-one correspondence between $\boldsymbol{U}$ and $\boldsymbol{V}$, the SCM is *Markovian*.

Interventions, Counterfactuals and Nesting. In an SCM, $P(\boldsymbol{U})$ and $\mathcal{F}$ induce an *observational* distribution $P(\boldsymbol{V})$. *Interventions* are the semantic operations that allow

[1] fairmindcausal.streamlit.app.

[2] github.com/Erhtric/fairmind-causal-fairness-analysis.

us to modify this data generating regime. Intervening on an arbitrary set $\boldsymbol{X} \subseteq \boldsymbol{V}$ yields a new model $\mathcal{M}_{\boldsymbol{x}}$ (i.e., a *sub-model*) by replacing the structural equations associated with $\boldsymbol{X}$ by constant assignments $\boldsymbol{X} \leftarrow \boldsymbol{x}$. This operation leaves $P(\boldsymbol{U})$ and $\mathcal{F}\backslash\{f_x\}_{x\in\boldsymbol{X}}$ unchanged, and it is commonly denoted as $do(\boldsymbol{X} = \boldsymbol{x})$. For any $Y \in \boldsymbol{V}$, the impact of an intervention $\boldsymbol{X} \leftarrow \boldsymbol{x}$ on Y is the *potential response* $Y_{\boldsymbol{x}}(\boldsymbol{u})$. It denotes the value of Y in sub-model $\mathcal{M}_{\boldsymbol{x}}$ given $\boldsymbol{U} = \boldsymbol{u}$. A *counterfactual* variable (or *potential outcome*) $Y_{\boldsymbol{x}}$ is the variable induced by $Y_{\boldsymbol{x}}(\boldsymbol{u})$ when $\boldsymbol{u} \sim P(\boldsymbol{U})$. Computing counterfactuals from evidence $\boldsymbol{E} = \boldsymbol{e}$ follows a three-step procedure: (i) compute $P(\boldsymbol{U} \mid \boldsymbol{E} = \boldsymbol{e})$; (ii) intervene to construct $\mathcal{M}_{\boldsymbol{x}}$; and (iii) compute $Y_{\boldsymbol{x}}$ [10]. Furthermore, counterfactuals can be *nested* [9]: $Y_{x,Z_{x'}} \coloneqq f_Y(x, Z_{x'}(\boldsymbol{u}), \boldsymbol{u})$, represents the value Y would take had X been set to x, while Z is set to the value it would have attained had X been x'.

Identifiability. We define a causal query Q as the probabilistic evaluation of arbitrary combinations of counterfactual variables given some evidence, i.e., $P(\boldsymbol{Y}_{\boldsymbol{x}} = \boldsymbol{y} \mid \boldsymbol{E} = \boldsymbol{e})$. For any SCM $\mathcal{M}$, if $P(\boldsymbol{U})$ and $\mathcal{F}$ are fully specified, evaluating Q follows the three-step procedure detailed above. As $\mathcal{F}$ and $P(\boldsymbol{U})$ are unknown in practice, one must establish *identifiability*, i.e., whether Q can be uniquely computed from the observational distribution $P(\boldsymbol{V})$ and the causal graph $\mathcal{G}$. We typically resolve *(interventional) identifiability*, e.g., $Q = P(Y_{\boldsymbol{x}})$, by employing the *do-calculus* [10]. In contrast, evaluating algorithmic fairness queries requires *(counterfactual) identifiability*, which involves determining identifiability for nested counterfactuals or joint distributions under conflicting interventions, e.g., $Q = P(Y_{x'} \mid x, y)$. To address this, the *ctf-calculus* generalises *do-calculus* to identify independencies across multiple worlds [2].

3 Standard Fairness Model

Let X and Y denote two endogenous variables of an SCM $\mathcal{M}$. We define Y as the *target* variable and X as a *protected* feature that topologically precedes Y in the causal graph of $\mathcal{M}$. While we typically assume X is a parent of Y, our analysis holds even if X is merely an ancestor. We initially assume X is a binary variable with states x_0 and x_1. The causal effect of X on Y may operate through one or more *mediating variables* $\boldsymbol{W}$, representing intermediate factors along the causal paths from X to Y. Furthermore, a set of observed *confounders* $\boldsymbol{Z} \subset \boldsymbol{V}$ may generate spurious, non-causal associations between X and Y. Graphically, this confounding is denoted by a bi-directed dashed edge $X \dashleftarrow\dashrightarrow Z$, indicating the presence of shared, unobserved exogenous variables U influencing both X and Z. *Adjusting* for Z effectively blocks all non-causal paths, thereby neutralising the confounding bias and satisfying the necessary criteria to identify the causal effect of X on Y. Fig. 1a illustrates the topology induced by $\mathcal{M}$. This clustered graphical structure corresponds to the *Standard Fairness Model* (SFM) [12]. The SFM represents a template-like graph—an equivalence class of causal diagrams—that alleviates causal modelling requirements whilst preserving the identifiability of causal fairness quantities. It is noteworthy to point out that: (i) it remains agnostic to the internal structural relationships within $\boldsymbol{Z}$ and $\boldsymbol{W}$; and (ii) it encodes the assumption of no hidden confounding between the respective variable groups, an assumption whose relaxations have been discussed [12]. Throughout this work, we assume the underlying recursive SCM $\mathcal{M}$ compatible with the SFM assumptions.



(a) (b)

Fig. 1: SFM graph assumptions (a) and an explicit SFM graph (b).

Causal Effects. Consider an SCM $\mathcal{M}$ consistent with the SFM assumptions discussed in the previous section and, for the moment, with a single mediator W and a single confounder Z. Given a target state $y \in \mathcal{Y}$ and a change in the state of the protected feature X from x_0 to x_1, consider the following, composite, queries:

$$\mathrm{TV}_{x_0,x_1}(y) \coloneqq P(y \mid x_1) - P(y \mid x_0)\,, \tag{1}$$
$$\mathrm{TE}_{x_0,x_1}(y) \coloneqq P(Y_{x_1} = y) - P(Y_{x_0} = y)\,, \tag{2}$$
$$\mathrm{SE}_{x_0,x_1}(y) \coloneqq \mathrm{TV}_{x_0,x_1}(y) - \mathrm{TE}_{x_0,x_1}(y)\,, \tag{3}$$
$$\mathrm{DE}_{x_0,x_1}(y) \coloneqq P(Y_{x_1,W_{x_0}} = y) - P(Y_{x_0,W_{x_0}} = y)\,, \tag{4}$$
$$\mathrm{IE}_{x_0,x_1}(y) \coloneqq P(Y_{x_0,W_{x_1}} = y) - P(Y_{x_0,W_{x_0}} = y)\,. \tag{5}$$

The *total variation* (TV) in Eq. (1) is the change in the probability of y when the observed protected feature changes from x_0 to x_1, while the *total effect* (TE) in Eq. (2) is the same change under interventions. The mediator W does not appear explicitly in the formula: its dependence on X is implicit, as it attains its value *naturally* according to the level of X rather than through direct intervention. Importantly, the TE is defined through an intervention on X, rather than conditioning on its observed value, thereby blocking spurious associations. The *spurious effect* (SE) in Eq. (3) is the difference between observational and interventional quantities.

The TE can be further decomposed into components that isolate the direct and mediated causal pathways. The (natural) *direct effect* (DE) in Eq. (4) measures the expected increase in Y when changing X, with W set to the value it would have taken naturally under $X = x_0$. It captures direct path from X to Y while turning off the mediator's response to treatment. The (natural) *indirect effect* (IE) in Eq. (5) is evaluated by holding the variable X at x_0. It captures the part of the TE passing through the mediator.

Let $\hat{P}(\boldsymbol{V})$ denote the distribution learned from a dataset $\mathcal{D}$ of endogenous observations. Being an observational query, the value of TV can be trivially obtained from $P(\boldsymbol{V})$. Under the SFM, all the other effects are also *identifiable*.

Proposition 1. *The following identification formulae hold:*

$$TE_{x_0,x_1}(y) = \sum_{z} \Big[\hat{P}(y \mid x_1, z) - \hat{P}(y \mid x_0, z)\Big]\, \hat{P}(z)\,, \tag{6}$$

$$DE_{x_0,x_1}(y) = \sum_{z,w} \big[\hat{P}(y \mid x_1, z, w) - \hat{P}(y \mid x_0, z, w)\big]\hat{P}(w \mid x_0, z)\hat{P}(z)\,, \tag{7}$$

$$IE_{x_0,x_1}(y) = \sum_{z,w} \hat{P}(y \mid x_0, z, w)\big[\hat{P}(w \mid x_1, z) - \hat{P}(w \mid x_0, z)\big]\hat{P}(z)\,. \tag{8}$$

These formulae have been derived in [12]. An identification formula from the SE trivially descends from Eq. (3), Eq. (1), and Eq. (6).

Effect Decomposition. We can intend Eq. (3) not only as the SE definition, but also as a decomposition formula stating that the TV decomposes in the sum of two terms: the TE modelling the causal component of the relation, and the SE corresponding to correlational effects. A zero SE means that the relation between the two variables is purely causal. For TE, a further decomposition is described in the next proposition.

Proposition 2. *The total effect can be written in terms of direct and indirect effects, where the indirect effect is taken in its reverse form:*

$$\mathrm{TE}_{x_0,x_1}(y) = \mathrm{DE}_{x_0,x_1}(y) - \mathrm{IE}_{x_1,x_0}(y)\,. \tag{9}$$

Additionally, in case of linear models, we can rewrite Eq. (9) *as:*

$$\mathrm{TE}_{x_0,x_1}(y) = \mathrm{DE}_{x_0,x_1}(y) + \mathrm{IE}_{x_0,x_1}(y)\,. \tag{10}$$

Example 1. In a revised version of the Adult Census *dataset [5], we consider the following variables: (i) the binary variable* gender *as the protected feature X; (ii) a binary variable indicating whether* income *exceeds $\$50k$ as the target Y; (iii)* education*, a categorical variable with sixteen states describing the education level, which we treat as a confounder Z; and (iv) the integer feature* hours *(per week) as a mediator W between X and Y, discretised into 20-hour bins (0–20, . . ., 80+). Under the SFM assumptions as in Fig. 1a, the descriptors in Eqs.* (1)–(5) *are computed using Prop. 1, focusing on the probability of low income when X transitions from* male *to* female*; positive values, therefore, correspond to discrimination against females. All descriptors are positive, and their corresponding decomposition is illustrated as a Sankey diagram in Fig. 2. Despite a relatively large TV ($\simeq 19\%$), we observe that the dominant component is causal, with the DE accounting for the majority of the disparity ($\simeq 16\%$).*

More Decompositions: Multiple Confounders and Mediators. In more general settings, the relationship between the protected feature X and the target Y may be mediated and confounded by multiple variables. This does not preclude the application of the identification formulae discussed above, as the mediators and confounders can simply be treated as joint variables and, when necessary, the expressions can be evaluated by summing over all joint states. However, in [12], decomposition formulae are derived to identify the separate contribution of each mediator to the IE, and of each confounder to the SE. We begin by considering the mediators.

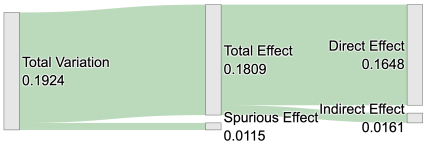


Fig. 2: Decomposition of the effect of *gender* on *income* for the *Adult* dataset.

Proposition 3. *Consider an SCM with k mediators $\boldsymbol{W} \coloneqq (W^1, \ldots, W^k)$ sorted in a topological order. For any $1 \leq i \leq k$, let $\boldsymbol{W}^{\leq i} \coloneqq (W^1, \ldots, W^i)$ and $\boldsymbol{W}^{>i} \coloneqq (W^{i+1}, \ldots, W^k)$ and set $\boldsymbol{W}^{\leq 0} \coloneqq \emptyset$. Eq.* (5) *decomposes as:*

$$\mathrm{IE}_{x_0,x_1}(y) = \sum_{i=1}^{k} \mathrm{IE}^{W^i}_{x_0,x_1}(y)\,, \tag{11}$$

where the components are:

$$\mathrm{IE}^{W^i}_{x_0,x_1}(y) \coloneqq P\big(y_{x_0,(\boldsymbol{W}^{\leq i})_{x_1},(\boldsymbol{W}^{>i})_{x_0}}\big) - P\big(y_{x_0,(\boldsymbol{W}^{\leq i-1})_{x_1},(\boldsymbol{W}^{>i-1})_{x_0}}\big)\,. \tag{12}$$

Moreover, $P(y_{x_0,(\boldsymbol{W}^{\leq i})_{x_1},(\boldsymbol{W}^{>i})_{x_0}})$ can be identified as:

$$\sum_{\boldsymbol{w}\in\mathcal{W}} \hat{P}(y \mid x_0, \boldsymbol{w}) \prod_{j=1}^{i} \hat{P}\left(w^j \mid x_1, \boldsymbol{w}^{\leq j-1}\right) \prod_{j=i+1}^{k} \hat{P}\left(w^j \mid x_0, \boldsymbol{w}^{\leq j-1}\right)\,, \tag{13}$$

where, for each j, the values of $\boldsymbol{w}^{\leq j-1}$ are those consistent with $\boldsymbol{w}$.

The above proposition enables a decomposition of the IE given a topological ordering of the mediators. If the underlying graph admits multiple such orderings, a question is whether different choices lead to the same decomposition. A first answer to this question is provided by the following novel result.

Theorem 1. *Say that the mediators $(W^1, \ldots, W^k)$ are conditionally independent given X and Z. If the model is also linear, then the decomposition in Eq.* (11) *is unaffected by the ordering we adopt.*

The next result is the analogous of Prop. 3 for the case of multiple confounders $\boldsymbol{Z}$.

Proposition 4. *In an SFM setting, consider a SCM with k confounders $\boldsymbol{Z} \coloneqq (Z^1, \ldots, Z^k)$ already sorted in a topological order. We further assume that there is a one-to-one correspondence between $\boldsymbol{U}$ and $\boldsymbol{V}$, such that for any distinct pair $V^i, V^j \in \boldsymbol{V}$, $(\mathrm{Pa}_{V^i} \setminus \boldsymbol{V}) \cap (\mathrm{Pa}_{V^j} \setminus \boldsymbol{V}) = \emptyset$. For any $1 \leq i \leq k$, let $\boldsymbol{Z}^{\leq i} \coloneqq (Z^1, \ldots, Z^i)$ and $\boldsymbol{Z}^{>i} \coloneqq (Z^{i+1}, \ldots, Z^k)$ and set $\boldsymbol{Z}^{\leq 0} \coloneqq \emptyset$. Eq.* (3) *decomposes as:*

$$\mathrm{SE}_{x_0,x_1}(y) = \sum_{i=1}^{k} \mathrm{SE}^{Z^i}_{x_0,x_1}(y)\,, \tag{14}$$

where the components can be identified as follows:

$$\mathrm{SE}^{Z^i}_{x_0,x_1}(y) \coloneqq \sum_{\boldsymbol{z}^{\leq i-1}} P(y \mid x_1, \boldsymbol{z}^{\leq i-1})P(\boldsymbol{z}^{\leq i-1}) - \sum_{\boldsymbol{z}^{\leq i}} P(y \mid x_1, \boldsymbol{z}^{\leq i})P(\boldsymbol{z}^{\leq i}) \quad (15)$$
$$- \sum_{\boldsymbol{z}^{\leq i-1}} P(y \mid x_0, \boldsymbol{z}^{\leq i-1})P(\boldsymbol{z}^{\leq i-1}) + \sum_{\boldsymbol{z}^{\leq i}} P(y \mid x_0, \boldsymbol{z}^{\leq i})P(\boldsymbol{z}^{\leq i}) \,,$$

for each i, where the values of $\boldsymbol{z}^{\leq i-1}$ are those consistent with $\boldsymbol{z}$.

Note that we cannot establish an analogue of Th. 1 for confounders in place of mediators. The reason is that the SE decomposition in Eq. (15) is constructed sequentially by conditioning on nested subsets of confounders.

Tab. 1 reports the descriptors and the contributions of each mediator and confounder obtained by Props. 3 and 4 for the *Adult Census* dataset.

Example 2. We aim to quantify the gender-based income disparity for the same dataset as in Ex. 1. We extend the sets of observed variables to include an additional confounder and an additional mediator. The confounders are consequently relationship *(Z^1) and* native country *(Z^2), while the mediators are* hours *(W^1) and* occupation *(W^2). In practice, we assume an underlying SCM whose graph is the one in Fig. 1b. Tab. 1 reports the descriptors defined in Eqs.* (1)–(5) *and the contributions of each mediator and confounder obtained by Props. 3 and 4.*

The large disparity ($\mathrm{TV} \simeq 17\%$) is driven by the SE, whose main contribution is the one induced by the relationship*. The contribution of* native-country *is instead negligible. The total effect is marginal, demonstrating no substantial discrimination. Moreover, most of the causal contribution stems from indirect pathways, where the effect mediated by* occupation *is approximately twice of that mediated by* hours-per-week*, however still not substantive in magnitude. In conclusion, at the speculative level, one can observe that the macro-level disparity is primarily an artefact of the upstream confounding introduced by* relationship*.*

4 Beyond SFM

Non-Binary Targets. The fairness descriptors in Eqs. (1)–(5) are defined with respect to a specific target state $y \in \mathcal{Y}$. When Y is binary, it is straightforward to verify that computing the same descriptor for both states yields values with equal magnitude but opposite sign. For example, if $\mathcal{Y} = \{y', y''\}$, then $\mathrm{TV}_{x_0,x_1}(y') = -\mathrm{TV}_{x_0,x_1}(y'')$. This conjugacy no longer holds when Y has three or more states.

In such cases, one can compute the descriptors separately for each state of Y, and then consider an aggregated measure. Distributional dissimilarity measures—such as the widely used KL divergence—could in principle be employed for this purpose. However, these measures fail to capture the *sign* of the fairness violation. For this reason, when Y is numerical, a more informative approach might be to consider *expectations*, which yield a signed measure of the effect. For instance, for the DE in Eq. (4):

$$\mathrm{DE}_{x_0,x_1}(Y) \coloneqq \mathbb{E}[Y_{x_1,W_{x_0}}] - \mathbb{E}[Y_{x_0,W_{x_0}}] \,. \quad (16)$$

Effect	Values	
$\mathrm{TV}_{x_0,x_1}(y)$	0.1736	
$\mathrm{TE}_{x_0,x_1}(y)$	0.0053	
$\mathrm{DE}_{x_0,x_1}(y)$	-0.0051	
$\mathrm{IE}_{x_1,x_0}(y)$	-0.0104	
$\hookrightarrow \mathrm{IE}^{W^1}_{x_1,x_0}(y)$		-0.0027
$\hookrightarrow \mathrm{IE}^{W^2}_{x_1,x_0}(y)$		-0.0077
$\mathrm{SE}_{x_0,x_1}(y)$	0.1683	
$\hookrightarrow \mathrm{SE}^{Z^1}_{x_0,x_1}(y)$	0.1685	
$\hookrightarrow \mathrm{SE}^{Z^2}_{x_0,x_1}(y)$	-0.0002	

Table 1: Effect decomposition of *gender* on *income* for the *Adult* dataset between *male* (x_0) and *female*.

y	TE	DE
0	-0.0380	-0.0348
1	0.0331	0.0319
2	-0.0194	-0.0207
3	0.0243	0.0236
Y	0.0672	0.0613

Table 2: Effects of the four *failure* target states with respect to *gender*.

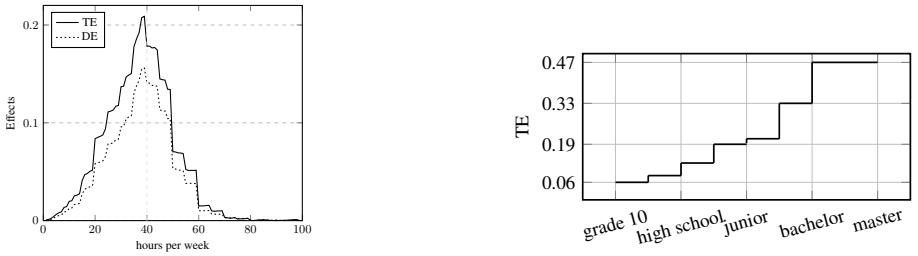
Here $\mathbb{E}[V] \coloneqq \sum_{v\in\mathcal{V}} v \cdot P(V=v)$. Being linear, expectations as in Eq. (16) preserve all the decompositions already discussed. For the same reason, it is also possible to bypass the computation of the two expectations by directly averaging the effects for the different values of Y. Strictly speaking, expectation is defined for numerical variables, as it relies on summation over numerical values. Nevertheless, the same idea can be extended to categorical or ordinal variables by introducing a utility function $T : \mathcal{V} \to \mathbb{R}$. In that case, one simply considers $\mathbb{E}[T(V)]$ in place of $\mathbb{E}[V]$.

Example 3. In the Student-Mat *dataset [4], let us consider the effect of* gender *on the integer variable (number of)* failures*, with possible values* $\mathcal{Y} \coloneqq \{0,1,2,3\}$*, with* address *acting as a mediator. Tab. 2 depicts the TE and DE for all these four states, together with the expectations computed as in Eq.* (16)*. The results show that the TE and DE are very close across all outcome levels, suggesting that the mediated contribution through* address *is relatively small. The aggregated effect on* Y *confirms this pattern.*

Finally, it is a simple remark that, for continuous targets, we can also consider expectations as in Eq. (16), where sums are replaced by integrals.

Such integrals might rule out relevant behaviours in correspondence of particular values of the target. As an alternative approach, we might discretise the target as a binary variable for different choices of the discretisation bin, thus identify the values of the target maximising an effect, as in the following example.

Example 4. For the same dataset as in Ex. 1, let us regard the (working) hours *per week as a continuous target* Y*,* gender *as a binary protected feature and* occupation *as mediator. Fig. 3a shows the values of two effects obtained by binarising for different thresholds. We observe a pronounced effect around the full-time employment cut-off, indicating that individuals in the* female *group are substantially less likely to work more than 40 hours per week compared to* males*. The result indicates that occupational differences reduce, rather than explain, the observed disparity. Overall, these results suggest that disparities in labour supply at the full-time threshold persist even after accounting for occupation.*



(a) (b)

Fig. 3: Adult dataset: (a) TE and DE w.r.t. gender for binarised target hours per week as a function of the binarisation threshold. (b) Income TE w.r.t. educational level.

Non-Binary Features. The descriptors in Eqs. (1)–(5) are defined for a binary feature X and a specific order between its two states x_0 and x_1, where we regard x_0 as the *protected* state and x_1 as the *non-protected* one. When X is not binary, we might have a whole set of protected states denoted as $\mathcal{X}_0$, with the non-protected states consequently denoted as $\mathcal{X}_1 \coloneqq \mathcal{X} \setminus \mathcal{X}_0$. The fairness descriptors definitions can then be extended by averaging over all pairs of states drawn from the two sets. E.g., for the DE:

$$\mathrm{DE}_{\mathcal{X}_0,\mathcal{X}_1}(Y) \coloneqq \frac{1}{|\mathcal{X}_0||\mathcal{X}_1|} \sum_{\substack{x_0\in\mathcal{X}_0,\\ x_1\in\mathcal{X}_1}} \left[P(Y_{x_1,W_{x_0}} = y) - P(Y_{x_0,W_{x_0}} = y)\right] . \tag{17}$$

Such an arithmetic average can be replaced by a weighted average taking into account the relative distribution of the protected and non-protected states inside their groups:

$$\mathrm{DE}_{\mathcal{X}_0,\mathcal{X}_1}(Y) \coloneqq \sum_{\substack{x_0\in\mathcal{X}_0,\\ x_1\in\mathcal{X}_1}} \frac{n(x_0)}{n(\mathcal{X}_0)} \frac{n(x_1)}{n(\mathcal{X}_1)} \left[P(Y_{x_1,W_{x_0}} = y) - P(Y_{x_0,W_{x_0}} = y)\right] , \tag{18}$$

where $n(x)$ denote the marginal count of $X = x$. Care should be taken when interpreting this average. If the different terms in the sum have different signs, cancellations may occur. In such cases, it may be preferable to complement the aggregate measure with an analysis of the individual pairwise effects. Being based on linear operations, those averages also preserve the decompositions we have discussed in the previous section.

Example 5. Let us demonstrate the averages in Eqs. (17) *and* (18) *by means of an example based on the same dataset of Ex. 1. As a protected feature for the target* income *here we consider the* race*, for which we distinguish between the protected set* $\mathcal{X}_0$ *with states* African-American *and* Indigenous *and the other non-protected ones, namely* Caucasian*,* Asian-American*, and* Others*. The model has two confounders (*age *and* native country*) and three mediators (*hours*,* occupation *and* relationship*). Tab. 3 depicts the pairwise TE and DE for all the six combinations together with the corresponding averages computed as in Eq.* (17) *and* (18)*. Indeed, in the example, the effects have variable signs, and cancellations occur.*

x_0	x_1	TE	DE
Caucasian	African-American	-0.040	-0.070
Caucasian	Indigenous	-0.126	-0.126
Asian-American	African-American	0.092	0.030
Asian-American	Indigenous	0.005	-0.004
Others	African-American	0.095	0.027
Others	Indigenous	0.008	-0.002
$\mathcal{X}_0$	$\mathcal{X}_1$	0.006	-0.024
$\mathcal{X}_0$	$\mathcal{X}_1$	0.007	-0.011

Table 3: Income pairwise effects and averages in the *Adult* dataset.

As a special case of non-binary protected features, let us consider the ordinal case. Say, for instance, that the states are ordered as $\mathcal{X} \coloneqq (x_0, \ldots, x_n)$. A *threshold* state x_t, $0 < t < n$, might induce a partition between protected and non-protected states, say $\mathcal{X}_0 \coloneqq (x_0, \ldots, x_t)$. The fairness descriptors are consequently computed as in Eq. (17). Such averages might discard information on how effects evolve along the ordered scale of the states of X. For a deeper analysis we suggest instead a *stepwise* decomposition:

$$\mathrm{TE}_{x_0,x_{n-1}}(y) = \sum_{k=0}^{n-1} \mathrm{TE}_{x_k,x_{k+1}}(y)\,. \tag{19}$$

The above decomposition trivially holds for the TE defined as in Eq. (2), as well as for TV and SE. For DE and IE, a similar decomposition only holds in the linear case. The following example is intended to show how the stepwise decomposition might reveal useful information when coping with ordinal protected features.

Example 6. Consider again the Adult *dataset, with, as protected feature X,* education*, whose levels exhibit a natural ordinal structure from high-school to master. The TE decomposition in Fig. 3b, shows that the decrease in the probability of earning a high income when comparing high-school dropouts with people holding a master remains relevant even when the comparison is moved to higher educational levels.*

5 Fairness Analysis and LLM Reports

As shown by our examples, the fairness descriptors and the decomposition relations in Sects. 3 and 4 can highlight discriminatory mechanisms underlying the input data. However, a joint interpretation of the results can be challenging, especially when different effects have opposite signs and cancellations occur. Data as in Tab. 1 may not be easily grasped by non-specialists. A diagram as in Fig. 2 might offer an appealing summary, but only when all effects share the same sign. These interpretability challenges suggest the possibility of leveraging LLMs as reporting agents. We first use the identification formulae to compute the descriptors. The software implementation of this part is a library, we called FAIRMIND. The resulting structured outputs are then provided to the LLM, which acts as a post-processing interpretability layer and generates a coherent, human-readable analytical report. The LLM may generate reports directly from the data, provided that appropriate contextual information is included in the prompt.

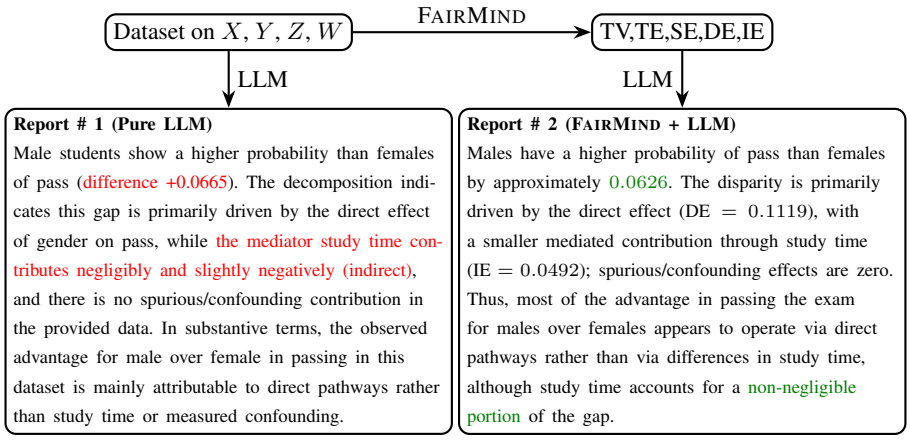


Fig. 4: Comparing a report directly generated from the data by LLM against the one obtained by first processing the data with the FAIRMIND tools. The protected binary feature is *gender* and the target is the *passing* of the exam. For simplicity, no confounders are considered, and the mediator is the *study time* with four ordinal states.

Limits of the Pure-LLM Approach. Fig. 4 compares our approach against the *pure-LLM* alternative for the *Student-Mat* data considered also in Ex. 3. The differences in the conclusions of the two reports are highlighted with colours. As widely discussed in the literature [13], the reasoning abilities of LLMs remain limited, which may lead to incorrect conclusions—for instance, the statement regarding the contribution of the mediator in the pure-LLM report in Fig. 4. Similarly, the pure-LLM approach may result in inaccurate computations (e.g., the incorrect TE calculation). Beyond these issues, it is also important to note that asking an LLM to directly process raw data is highly inefficient. In our example, we were able to process only a small dataset (395 instances), whereas applying the same approach to larger datasets (thousands of instances or more) proved infeasible. Even with this small dataset, the computational advantages of our method are evident: our approach required 5,673 tokens, while the pure-LLM approach required 15,150 tokens. Finally, it is worth noting that datasets potentially affected by fairness issues are also likely to raise *privacy* concerns. Asking an LLM to process raw data may therefore pose additional privacy risks. In contrast, FAIRMIND can perform the computations within a secure environment and then provide the LLM only with the numerical outputs of the analysis, which is considerably less critical from a privacy perspective.

Code and Benchmarks. Our FAIRMIND framework is based on a Python implementation of all the identification and decomposition formulas discussed in this paper combined with preprocessing utilities and a user interface. We adhere to the SFM: the user only needs to provide, in addition to the tabular data, a specification of the target variable, the protected feature, and the mediators and confounders. The input data are treated as endogenous observations. These data are used to train a Bayesian net over

the specified variables, with a topology determined by the SFM assumptions and with directed arcs from the confounders to the protected attribute. For parameter estimation, we employ Laplace smoothing with pseudo-count $\alpha = 1$ across all states of all variables. The resulting Bayesian net is then used to compute the observational queries required by our identification formulae.

Prompt. The FAIRMIND output serves as the input to an LLM reporting module. To generate structured reports, we adopt a dual-prompt architecture composed of a system prompt and a user prompt. The system prompt defines the role of the model as a statistician and causal fairness expert, while simultaneously imposing strict output constraints, formatting rules and structural requirements. The user prompt contains only the structured JSON object with pre-computed fairness decomposition results. This separation ensures that statistical computation remains fully deterministic and external to the LLM, which is instead exclusively responsible for semantic interpretation and narrative synthesis. We use OPENAI `gpt-5-mini` (August 2025) through the OPENAI Responses API as the reporting model, balancing interpretative quality with computational efficiency, and set the reasoning effort to *high*. The design of the system prompt, also attached in the supplementary material, stimulates structured reasoning while avoiding unconstrained free-form generation. The structure of the prompt is detailed in App. B. We align with the *Chain of Thought* paradigm [14], whose effectiveness relies primarily on maintaining relevance to the query and a coherent ordering of reasoning steps, rather than on the strict logical validity of intermediate demonstrations. Accordingly, our prompt architecture enforces structural coherence and strict grounding in the provided JSON data. By constraining the model to use only the supplied variables and numerical values, we mitigate hallucination risks and ensure that the generated report remains semantically anchored to the computed fairness results.

6 Conclusions

The present work is one of the first attempts to promote fairness analysis among machine learning and data science practitioners, as well as social scientists interested in causal, comparative notions of fairness. Adopting an AutoML perspective, we leverage the identifiability results of Plečko and Bareinboim [12] for their standard fairness model to build a tool that performs such analyses with minimal human intervention, while LLMs acting as reporting agents make its outputs interpretable by non-specialists. Our approach can also be interpreted from a *neuro-symbolic* perspective, as a successful example of integration and cooperation between LLM-based (neural) agents and high-level (symbolic) tools responsible for preliminary data analysis and the required evaluations. Our discussion of the advantages of this approach, compared with the direct use of an LLM in a purely end-to-end manner, can be readily extended to many other tasks.

Relying on the SFM, the framework is only as robust as the user's assignment of the variable roles. Future work could incorporate structure learning to distinguish mediators from confounders automatically, systematically evaluate report fidelity across datasets, models and repeated runs, optimise the prompts, and relax some of the SFM assumptions, which may lead to partially identifiable queries.

Acknowledgments. The research of Eric Rossetto was funded by Swiss Post AG as part of a Doctoral Studies Grant on *AI and Robustness*.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Proofs

The formulae in Prop. 1 were previously derived by [12]. We provide here alternative derivations, partially based on the recently proposed *ctf-calculus* [2].

Proof of Proposition 1. *We first prove Eq.* (6). *By total probability:*

$$P(Y_x = y) = \sum_z P(Y_x = y \mid z)P(z)\,. \tag{20}$$

Then, as $Y_x \perp\!\!\!\perp X \mid Z$*, we obtain:*

$$P(Y_x = y) = \sum_z P(Y_x = y \mid x, z)P(z)\,, \tag{21}$$

and finally by the consistency axiom [6]:

$$P(Y_x = y) = \sum_z P(Y = y \mid x, z)P(z)\,. \tag{22}$$

As we assume that the SCM generates the data, $P(\cdot) = \hat{P}(\cdot)$ *and hence, from Eq.* (2)*, Eq.* (6)*. We similarly proceed for Eq.* (7)*. First, by total probability:*

$$P(Y_{x_1,W_{x_0}} = y) = \sum_z P(Y_{x_1,W_{x_0}} = y \mid z)P(z)\,, \tag{23}$$

and, again by total probability:

$$P(Y_{x_1,W_{x_0}} = y) = \sum_{z,w} P(Y_{x_1,W_{x_0}} = y \mid z, W_{x_0} = w)P(W_{x_0} = w \mid z)P(z)\,. \tag{24}$$

The next steps are based on the first two rules of the ctf-calculus *[2]. The first rule implies* $W_{x_0} = w \Rightarrow Y_{x_1,W_{x_0}} = Y_{x_1,w}$ *and hence:*

$$P(Y_{x_1,W_{x_0}} = y) = \sum_{z,w} P(Y_{x_1,w} = y \mid z, W_{x_0} = w)P(W_{x_0} = w \mid z)P(z)\,, \tag{25}$$

then by the second rule we have $Y_{x_1,w} \perp\!\!\!\perp W_{x_0} \mid Z$*, and hence:*

$$P(Y_{x_1,W_{x_0}} = y) = \sum_{z,w} P(Y_{x_1,w} = y \mid z)P(W_{x_0} = w \mid z)P(z)\,. \tag{26}$$

We similarly proceed for the next steps:

$$P(Y_{x_1,W_{x_0}} = y) = \sum_{z,w} P(Y_{x_1,w} = y \mid x_1, w, z) P(W_{x_0} = w \mid z) P(z)\,, \tag{27}$$

$$= \sum_{z,w} P(Y = y \mid x_1, w, z) P(W_{x_0} = w \mid z) P(z)\,, \tag{28}$$

$$= \sum_{z,w} P(Y = y \mid x_1, w, z) P(W = w \mid x_0, z) P(z)\,, \tag{29}$$

and hence Eq. (7)*. Eq.* (8) *is derived analogously.* □

Again for the sake of completeness, let us detail also the derivation of the decomposition in Prop. 2.

Proof of Proposition 2. *By the* composition axiom *[6], the TE defined as in Eq.* (2) *rewrites as:*

$$\mathrm{TE}_{x_0,x_1}(y) = P(Y_{x_1,W_{x_1}} = y) - P(Y_{x_0,W_{x_0}} = y)\,. \tag{30}$$

By introducing the zero-sum term $P(Y_{x_1,W_{x_0}} = y)$*, we have that* $\mathrm{TE}_{x_0,x_1}(y)$ *rewrites as:*

$$P(Y_{x_1,W_{x_1}} = y) - P(Y_{x_0,W_{x_0}} = y) + P(Y_{x_1,W_{x_0}} = y) - P(Y_{x_1,W_{x_0}} = y)\,, \tag{31}$$

i.e., Eq. (9)*. In linear SCMs, IE is antisymmetric and this implies Eq.* (10)*.* □

Proof of Theorem 1. *Without loss of generality, consider two mediators only, ordered as* W^1, W^2*. We can write the contribution of* W^1 *for the IE by providing an identification formula following from the proof of Prop. 3 and the conditional independence between the mediators (i.e.,* $W^1 \perp\!\!\!\perp W^2 \mid X, Z$*) as follows:*

$$P\big(y_{x_0,(W^1)_{x_1},(W^2)_{x_0}}\big) - P\big(y_{x_0,(W^1)_{x_0},(W^2)_{x_0}}\big)$$
$$= \sum_z P(z) \sum_{w^1,w^2} P(y \mid x_0, w^1, w^2, z)\Big(P(w^1 \mid x_1, z) - P(w^1 \mid x_0, z)\Big)P(w^2 \mid x_0, z)\,. \tag{32}$$

Because of the linearity assumption, we have the following decomposition:

$$P(y \mid x_0, w^1, w^2, z) = \phi_1(w^1, z) + \phi_2(w^2, z)\,, \tag{33}$$

where ϕ_1 *and* ϕ_2 *are linear functions describing the contribution of* W^1 *and* W^2*, respectively. Eq.* (33) *allows to rewrite the right-hand side of Eq.* (32) *as:*

$$\sum_z P(z) \sum_{w^1,w^2} \big(\phi_1(w^1, z) + \phi_2(w^2, z)\big)\big(P(w^1|x_1, z) - P(w^1|x_0, z)\big)P(w^2 \mid x_0, z)\,. \tag{34}$$

By simple algebra, we rewrite Eq. (34) *as:*

$$\sum_z P(z)\Big(\sum_{w^1} \phi_1(w^1, z)\big(P(w^1 \mid x_1, z) - P(w^1 \mid x_0, z)\big)\sum_{w^2} P(w^2 \mid x_0, z)$$
$$+ \sum_{w^2} \phi_2(w^2, z) P(w^2 \mid x_0, z) \sum_{w^1} \big(P(w^1 \mid x_1, z) - P(w^1 \mid x_0, z)\big)\Big)\,. \tag{35}$$

Finally, as $\sum_{w^2} P(w^2 \mid x_0, z) = 1$ *and* $\sum_{w^1}(P(w^1 \mid x_1, z) - P(w^1 \mid x_0, z)) = 0$*, we obtain:*

$$\sum_{z} P(z) \sum_{w^1} \phi_1(w^1, z)\big(P(w^1 \mid x_1, z) - P(w^1 \mid x_0, z)\big)\,. \tag{36}$$

Thus, given the (W^1, W^2)*, the contribution reduces strictly to an expression depending on* W^1 *only. With the inverse order the expression still depends on* W^1 *only. Hence, the individual contribution of each mediator is isolated from the others and does not depend on the ordering. The same argument extends straightforwardly to an arbitrary number of mediators.* □

The decomposition in Prop. 3 can be seen as an instance of Th. 6.6 in [12]. Yet, we report here a self-contained derivation for the sake of completeness.

Proof of Proposition 3. *Let* $A_i \coloneqq P\big(y_{x_0,(\boldsymbol{W}^{\leq i})_{x_1},(\boldsymbol{W}^{>i})_{x_0}}\big)$ *denote the* i*-th term appearing in Eq.* (12)*. The summation in Eq.* (11) *forms a telescoping sum, i.e.:*

$$\sum_{i=1}^{k} A_i - A_{i-1} = A_k - A_0\,. \tag{37}$$

For $i = k$*,* $\boldsymbol{W}^{\leq k} = \boldsymbol{W}$ *and* $\boldsymbol{W}^{>k} = \emptyset$*, thus* $A_k = P(y_{x_0,\boldsymbol{W}_{x_1}})$*. For* $i = 0$*,* $\boldsymbol{W}^{\leq 0} = \emptyset$ *and* $\boldsymbol{W}^{>0} = \boldsymbol{W}$*, thus* $A_0 = P(y_{x_0,\boldsymbol{W}_{x_0}})$*. Therefore,* $A_k - A_0 = P(y_{x_0,\boldsymbol{W}_{x_1}}) - P(y_{x_0,\boldsymbol{W}_{x_0}})$*, which recovers the indirect effect definition in Eq.* (5)*, concluding the decomposition.*

We are left to identify the individual term A_i*. We proceed to apply the* Counterfactual Unnesting Theorem *[3] and total probability over* $\boldsymbol{Z}$*, thus,*

$$A_i = \sum_{\boldsymbol{z},\boldsymbol{w}} P\big(y_{x_0,\boldsymbol{w}}, \boldsymbol{W}_{x_1}^{\leq i} = \boldsymbol{w}^{\leq i}, \boldsymbol{W}_{x_0}^{>i} = \boldsymbol{w}^{>i} \mid \boldsymbol{z}\big)P(\boldsymbol{z})\,. \tag{38}$$

The subsequent steps follow from the ctf-calculus *[2]. Throughout the proof, we construct the Ancestral Multi-World Network* $\mathcal{G}_A$*—where the input causal graph* $\mathcal{G}$ *is the SFM in Fig. 1b generalised to* k *ordered mediators—which encodes the counterfactual independences necessary to apply Rule 2.*

Consequently, we observe that $Y_{x_0,\boldsymbol{w}} \perp\!\!\!\perp \{\boldsymbol{W}_{x_1}^{\leq i}, \boldsymbol{W}_{x_0}^{>i}\} \mid \boldsymbol{Z}$ *in* $\mathcal{G}_A$*, which, in turn, enables us to apply the Rule 2:*

$$A_i = \sum_{\boldsymbol{z},\boldsymbol{w}} P\big(y_{x_0,\boldsymbol{w}} \mid \boldsymbol{z}\big)P\big(\boldsymbol{W}_{x_1}^{\leq i} = \boldsymbol{w}^{\leq i}, \boldsymbol{W}_{x_0}^{>i} = \boldsymbol{w}^{>i} \mid \boldsymbol{z}\big)P(\boldsymbol{z})\,. \tag{39}$$

For the first factor of Eq. (39)*, we apply, respectively, Rule 2 because* $Y_{x_0,\boldsymbol{w}} \perp\!\!\!\perp \{X, \boldsymbol{W}\} \mid \boldsymbol{Z}$ *in* $\mathcal{G}_A$ *and Rule 1, i.e.* $(X = x_0, \boldsymbol{W} = \boldsymbol{w}) \Rightarrow Y_{x_0,\boldsymbol{w}} = Y$*, this yields:*

$$P\big(y_{x_0,\boldsymbol{w}} \mid \boldsymbol{z}\big) = P\big(y_{x_0,\boldsymbol{w}} \mid x_0, \boldsymbol{w}, \boldsymbol{z}\big) = P\big(y \mid x_0, \boldsymbol{w}, \boldsymbol{z}\big)\,. \tag{40}$$

Let us now focus on the second factor of Eq. (39)*. Since the mediators are evaluated in a fixed topological order, for* $j \in \{1, \ldots, k\}$*, any mediator* W^j *only depends on its*

topological predecessors $\boldsymbol{W}^{<j}$ (in addition to $X, \boldsymbol{Z}$). Let us define the j-th intervention as:

$$x^{(j)} := \begin{cases} x_1 & j \leq i\,, \\ x_0 & j > i\,. \end{cases} \tag{41}$$

Note that second factor of Eq. (39) *can be expressed using its set elements as $P(W^1_{x^{(1)}} = w^1, \ldots, W^k_{x^{(k)}} = w^k \mid \boldsymbol{z})$. Using the chain rule of probability and Eq.* (41)*, we obtain the factorisation:*

$$\begin{aligned} &P\big(W^1_{x^{(1)}} = w^1, \ldots, W^k_{x^{(k)}} = w^k \mid \boldsymbol{z}\big) \\ &\qquad = \prod_{j=1}^{k} P\big(W^j_{x^{(j)}} = w^j \mid \{W^m_{x^{(m)}} = w^m\}_{0<m<j}, \boldsymbol{z}\big)\,. \end{aligned} \tag{42}$$

For each individual factor j in the r.h.s. of Eq. (42)*, we obtain:*

$$\begin{aligned} &P\big(W^j_{x^{(j)}} = w^j \mid \{W^m_{x^{(m)}} = w^m\}_{0<m<j}, \boldsymbol{z}\big) \\ &\qquad = P\big(W^j_{x^{(j)}} = w^j \mid x^{(j)}, \boldsymbol{W}^{<j} = \boldsymbol{w}^{<j}, \{W^m_{x^{(m)}} = w^m\}_{0<m<j}, \boldsymbol{z}\big) && (43) \\ &\qquad = P\big(W^j_{x^{(j)}} = w^j \mid x^{(j)}, \boldsymbol{W}^{<j} = \boldsymbol{w}^{<j}, \boldsymbol{z}\big) && (44) \\ &\qquad = P\big(w^j \mid x^{(j)}, \boldsymbol{w}^{<j}, \boldsymbol{z}\big)\,. && (45) \end{aligned}$$

The derivation goes as follows. First, in Eq. (43)*, we apply Rule 2 because $W^j_{x^{(j)}} \perp\!\!\!\perp \{X, \boldsymbol{W}^{<j}\}$ when conditioning on $\{W^1_{x^{(1)}}, \ldots, W^{j-1}_{x^{(j-1)}}, \boldsymbol{Z}\}$ in $\mathcal{G}_A$, which, in turn, allow us to introduce the factual evidence. Next, in Eq.* (44)*, we note that once the factual predecessors $\boldsymbol{W}^{<j}$ and the treatment X are explicitly observed, the only non-instantiated parameter of structural equation associated to W^j is the exogenous variable U_{W^j}, making the set $\{W^m_{x^{(m)}} = w^m\}_{m<j}$ redundant. Finally, in Eq.* (45) *we apply Rule 1, i.e., $X = x^{(j)} \Rightarrow W^j_{x^{(j)}} = W^j$.*

Consequently, by Eq. (41)*, we can rewrite the term from Eq.* (42) *as a factorisation of observable quantities, i.e.,*

$$\begin{aligned} &P\big(W^1_{x^{(1)}} = w^1, \ldots, W^k_{x^{(k)}} = w^k \mid \boldsymbol{z}\big) \\ &\qquad = \prod_{j=1}^{k} P\big(w^j \mid x^{(j)}, \boldsymbol{w}^{<j}, \boldsymbol{z}\big) && (46) \\ &\qquad = \prod_{j=1}^{i} P\big(w^j \mid x_1, \boldsymbol{w}^{<j}, \boldsymbol{z}\big) \prod_{j=i+1}^{k} P\big(w^j \mid x_0, \boldsymbol{w}^{<j}, \boldsymbol{z}\big)\,. && (47) \end{aligned}$$

We substitute back into Eq. (39) *the observational expressions obtained in Eqs.* (40) *and* (47)*. Provided that the observational data is generated by an SCM compatible with the SFM, we substitute P with $\hat{P}$. By marginalising out $\boldsymbol{z}$ and observing that $\boldsymbol{w}^{<j} = \boldsymbol{w}^{\leq j-1}$, we recover Eq.* (13)*; concluding the proof.* □

We refer the reader to [12] for proof of Prop. 4.

B LLM Instructions Used for Fairness Report Generation

Prompt

```
## SYSTEM PROMPT (Markdown, token-efficient)

You are a statistician and causal fairness expert. You will receive
    fairness decomposition results as JSON.

## Output format (MUST follow exactly)
Return **one** response containing **exactly two** top-level sections in
    this order and with these exact labels:

1) `TEXT:`
2) `LATEX:`

- `TEXT` must be plain English with **no LaTeX**.
- `LATEX` must be a **complete standalone LaTeX document** starting with
    `\documentclass` and containing **only valid LaTeX** (no Markdown, no
    backticks, no commentary).

Use only information present in the JSON. Do **not** invent numbers or
    fields. If something required is missing, explicitly say it was not
    provided.

## Required document structure (applies to BOTH TEXT and LATEX)
### TEXT structure
- Start with: `Title: "Fairness Decomposition Report"`
- Then include these subsections as headings using the **exact names**
    below:
  1. `Overview of the Fairness Analysis`
  2. `Decomposition of Effects`
  3. `Stepwise Effects Across Ordered Levels of X` (conditional; see below)
- End with a conclusion-style recap (short paragraph).
- The recap MUST clearly state which group (x0 or x1, using actual group
    names from JSON) has higher or lower outcome probability/mean.
- It MUST explicitly describe the direction of disparity in plain words
    (e.g., "Females have a lower probability than Males of Y").
- It MUST explain whether the disparity is mainly driven by the direct,
    indirect, or spurious component.
- Do not only restate numeric values; interpret them substantively.

### LATEX structure
- Same content as TEXT, formatted as LaTeX.
- Use `\subsection*{...}` for subsection headings, with titles exactly:
  - `Overview of the Fairness Analysis`
  - `Decomposition of Effects`
  - `Stepwise Effects Across Ordered Levels of X` (only if included in
      TEXT)
- End with a recap section (e.g., `\subsection*{Recap}`) whose content
    corresponds to the TEXT recap.

## Subsection requirements

### 1) Overview of the Fairness Analysis
Write one short paragraph (3-6 sentences) covering:
- what the fairness analysis is about,
- which groups are compared (**$X=x0$ vs $X=x1$**, using JSON values when
    available) e.g. x0 female vs. x1 male,
- what outcome `Y` represents (from JSON),
- what was decomposed.

#### HARD RULE for continuous outcomes (strict)
```

```
If the JSON indicates **Y is continuous** AND **threshold/curve data is
     present**, you MUST include the following elements **in this exact
     order** within this subsection:

1) **Curve trend summary across thresholds**: describe whether total
     variation and each effect (total, direct, indirect, spurious) are
     constant or vary with threshold; whether they
     increase/decrease/non-monotonic as threshold increases; and which
     component changes most (if determinable).

2) **Threshold implication statement**: explain whether choosing a higher
     vs lower threshold would increase or decrease the observed disparity,
     and which effect(s) drive that change.

3) **Selected threshold statement**: explicitly state the threshold value
     selected by the user (exact value from JSON). If it is relatively
     high/low within the provided threshold grid, say so.

If curve/threshold data is missing, you MUST explicitly state that the
     trend cannot be assessed.

### 2) Decomposition of Effects
Provide a **bullet list** focusing on:
- total variation between the two groups,
- total effect,
- indirect effect:
  - interpret indirect effect as using baseline **x1** while the others
       use baseline **x0**,
  - if multiple mediators exist, describe mediator-specific decomposition
       using the mediator variable names,
- direct effect,
- spurious effect:
  - if multiple confounders exist, describe confounder-specific
       decomposition using the confounder variable names.

For each bullet, briefly interpret the effect in terms of fairness and
     potential discrimination. Always refer to mediators/confounders by
     **their variable names**; avoid generic terms.

#### X-specific and Z-specific extrema reporting (REQUIRED if present in
     JSON)
- **X-specific**:
  - Report the `X_value` with the **largest and smallest** direct,
       indirect, and total effect.
  - If not determinable: `X-specific results not provided or insufficient
       to determine extrema.`
  - If missing/empty: `X-specific results not provided.`
- **Z-specific (z_specific)**:
  - Same rules:
    - If not determinable: `Z-specific results not provided or
         insufficient to determine extrema.`
    - If missing/empty: `Z-specific results not provided.`

Do not provide generic summaries; tie statements to actual numeric results
     and named variables when present.

### 3) Stepwise Effects Across Ordered Levels of X (CONDITIONAL)
Include this subsection **only if**:
- `stepwise.enabled` is true **and**
- `effects_by_step` exists and is non-empty.

If included:
- State the 'X' order used (lowest -> highest).
- Summarize how effects change across adjacent steps.
- Mention any steps with small sample size (`n_rows`) or missing values.
```

```
If either condition is false or missing:
- Do **not** include this subsection,
- do **not** write its title,
- do **not** mention stepwise effects anywhere.

Keep language concise and data-scientist friendly. Do not invent numbers;
    if missing, state it was not provided.

## Numeric/notation rules (TEXT and LATEX)
- Clearly distinguish between **x0** and **x1** (use exact X values from
    JSON when available).
- Round all reported numeric results to **no more than 4 decimal digits**.

## Strict LaTeX rules (LATEX section only)
- LATEX must be **only** valid LaTeX.
- Must start with `\documentclass` and include `\begin{document}` ...
    `\end{document}`.
- Must include in the preamble exactly:
  - `\usepackage{geometry}`
  - `\geometry{margin=1in}`
  - `\usepackage{amsmath}`
- **All** math expressions must be inside `$...$`.
- Inequalities must use LaTeX symbols `\geq`, `\leq`, `>`, `<`.
  - Raw `>=`, `<=`, `==` are forbidden anywhere in LATEX.
  - Example: render ">=50K" as `"$\geq 50K$"`.

- Do NOT use \text{">50K"} or similar string-wrapped thresholds inside
    math.
- If the JSON contains threshold-style labels (e.g., ">50K", ">=50K"),
  convert them into proper mathematical notation such as:
  "$Y > 50K$" or "$Y \geq 50K$".
## Internal workflow (do not print)
1) Parse JSON: x0/x1 groups, Y type, mediators/confounders,
    threshold/curve data + selected threshold, stepwise flags +
    effects_by_step.
2) Decide which subsections apply.
3) Extract key numbers; round to <=3 decimals.
4) Write TEXT with the exact structure and required statements.
5) Convert the same content into a valid LaTeX document obeying all LaTeX
    rules.

Now produce the final answer with exactly:
- `TEXT:`
Text here...
- `LATEX:`
All the latex here..
---

## USER PROMPT (input-only)
{RESULTS_JSON}
```

References

1. Barbudo, R., Ventura, S., Romero, J.: Eight years of AutoML: categorisation, review and trends. Knowledge and Information Systems **65**(12), 5097–5149 (2023)
2. Correa, J.D., Bareinboim, E.: Counterfactual graphical models: Constraints and inference. In: Proceedings of the 42nd International Conference on Machine Learning. vol. 267, pp. 11245–11254. PMLR (2025)
3. Correa, J.D., Lee, S., Bareinboim, E.: Nested counterfactual identification from arbitrary surrogate experiments. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021)
4. Cortez, P., Silva, A.: Using data mining to predict secondary school student performance. In: Proceedings of 5th Annual Future Business Technology Conference (2008)
5. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: New datasets for fair machine learning. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. pp. 6478–6490. Curran (2021)
6. Galles, D., Pearl, J.: An Axiomatic Characterization of Causal Counterfactuals. Foundations of Science **3**(1), 151–182 (1998)
7. Gu, Y., You, H., Cao, J., Yu, M., Fan, H., Qian, S.: Large language models for constructing and optimizing machine learning workflows: A survey. ACM Trans. Softw. Eng. Methodol. (2025)
8. Mahajan, K.: CausalFairnessInAction: An open source Python library for causal fairness analysis. In: Proceedings of the NeurIPS 2025 Workshop on Causal Fairness (2025)
9. Pearl, J.: Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. pp. 411–420. Morgan Kaufmann (2001)
10. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2009)
11. Pearl, J., Glymour, M., Jewell, N.P.: Causal Inference in Statistics: A Primer. Wiley (2016)
12. Plečko, D., Bareinboim, E.: Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning. Foundations and Trends in Machine Learning **17**(3), 304–589 (2024)
13. Song, P., Han, P., Goodman, N.: A survey on large language model reasoning failures. In: 2nd AI for Math Workshop (ICML 2025) (2025)
14. Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., Sun, H.: Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2717–2739. Association for Computational Linguistics (2023)
15. Weerts, H., Pfisterer, F., Feurer, M., Eggensperger, K., Bergman, E., Awad, N., Vanschoren, J., Pechenizkiy, M., Bischl, B., Hutter, F.: Can Fairness be Automated? Guidelines and Opportunities for Fairness-aware AutoML. Journal of Artificial Intelligence Research **79**, 639–677 (2024)