FAIRMIND: A Tool for Automatic Causal Fairness Analysis with LLM-Generated Reports

ALESSIA BERARDUCCI, ERIC ROSSETTO, ALESSANDRO ANTONUCCI, and MARCO ZAFFALON, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), USI-SUPSI, Switzerland

We introduce FAIRMIND, a software tool designed to automate fairness analysis at the dataset level. Our approach builds on the assumptions of the *standard fairness model* recently proposed by Plečko and Bareinboim, enabling a principled evaluation of fairness in terms of causal effects. The analysis relies on *counterfactual* queries involving the target variable, potential mediators and confounders and different values of an input feature treated as *protected*. Following the necessary data preprocessing, the tool performs closed-form computations of these effects. Large language models (LLMs) are then employed to generate detailed and interpretable reports on the fairness properties identified in the training data. Even in zero-shot settings, our approach offers substantial advantages over direct LLM-based analyses, particularly as the complexity of the data structures increases.

Reference Format:
Alessia Berarducci, Eric Rossetto, Alessandro Antonucci, and Marco Zaffalon. 2026. FAIRMIND: A Tool for Automatic Causal Fairness Analysis with LLM-Generated Reports. In *Proceedings of Fifth European Conference on Algorithmic Fairness (ECAF'26).* Proceedings of Machine Learning Research, 6 pages.

1 Introduction

Algorithmic Fairness (AF) field is about uncovering and rectifying unjustified bias along protected attributes, e.g., race, gender, age, class; as such biases emerge from statistical and machine-learning models employed thoroughly in automated decision-making [2, 10]. Traditionally, AF promises have been addressed by a plethora of correlative approaches based on statistical analysis or machine learning concepts [4, 6]. This might capture disparities across groups, but typically does not account for the causal mechanisms generating these differences. To overcome these limitations, recent scholarship has argued for extending equitable approaches from a causal perspective [5, 11, 16], which becomes a necessity for complex sociological variables like ethnicity [15]. Specifically, Plečko and Bareinboim [14] extend tools from mediation analysis to derive closed-form AF *identifiable* expressions under specific structural assumptions. This enables a decomposition of discriminatory effects along distinct causal pathways. However, practical adoption of causal fairness analysis remains limited. To the best of our knowledge, only one tool explicitly implements causal methods [8]; however, it is restricted to aggregated effects and is not publicly available. Second, the resulting causal fairness descriptors are highly technical, hindering correct interpretation by non-specialists. On a related note, recent work has shown that LLMs can support the

Authors' Contact Information: Alessia Berarducci; Eric Rossetto; Alessandro Antonucci; Marco Zaffalon, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), USI-SUPSI, Lugano, Switzerland, alessia.berarducci@supsi.ch, eric.rossetto@supsi.ch, alessandro.antonucci@supsi.ch, marco.zaffalon@supsi.ch.

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

ECAF'26, September 02–September 04, 2026, Ghent, BE

© 2026 Copyright held by the owner/author(s).

communication of complex model explanations by transforming technical outputs into accessible natural-language narratives [9]. This motivates the use of LLMs as an interpretative layer for formally computed causal fairness metrics. To address these challenges, we present here FAIRMIND,[1] a software tool to compute the causal AF metrics and feed them into an LLM to generate detailed and accessible AF textual reports.

2 Standard Fairness Model (SFM)

SCMs and SFM. Variable V takes its states from $\mathscr{V}$ and $v \in \mathscr{V}$ is a generic state. In a *structural causal model* (SCM) [13], a set of so-called *endogenous* variables $\boldsymbol{V}$ is in one-to-one correspondence with the nodes of a directed *causal* graph $\mathscr{G}$ describing the underlying causal relations. Directed edges represent causal relationships and bi-directed dotted edges denote latent confounding effects. The state of each $V \in \boldsymbol{V}$ is determined by structural equations whose inputs are the parents of V and an additional, exogenous, variable. Confounding effects are induced when two variables share the same exogenous variable. The SFM corresponds to the topology in Fig. 1 [14], where the effect of a protected attribute X on a target variable Y can be direct or mediated by W and confounded by Z. Exogenous variables are typically latent, i.e., unobservable. We assume only a dataset of endogenous observations generated by the SCM to be available, from which we learn a joint distribution $P(\boldsymbol{V})$ to compute any *observational* query, i.e., the probability of a state of an endogenous variable, possibly given a state of one or many other variables. As a (toy) example of an SFM setup, we consider the AF of an automated banking algorithm to determine the credit card limits based on four Boolean variables. The protected attribute X indicates whether the age of the applicant is above a given threshold. The system includes a wealth-tracking confounder Z that reports whether the liable assets exceed a specific threshold and a credit score mediator W denotes problematic (i.e., bad or good) credit histories. The outcome Y indicates whether the assigned credit limit exceeds a threshold. Simulated data are in Tab. 1.

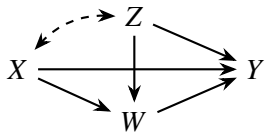


Fig. 1. SFM causal graph.

Table 1. Counts for the toy example.

		(W,Y)			
Z	X	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
0	0	242	31	15	17
0	1	87	31	21	89
1	0	87	42	10	57
1	1	32	24	0	215

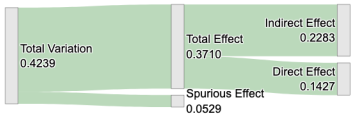


Fig. 2. Effects decomposition.

Interventions & Counterfactuals. An *intervention* alters an SCM by replacing the structural equations of a subset of endogenous variables with constant assignments. Such operations yield a new SCM, with the same structural equations but the intervened ones. We denote the variables of such post-interventional SCM by subscripts, e.g., after an intervention forcing $X = x_0$ in our running example, the target and the mediator can be rewritten as Y_{x_0} and W_{x_0}. *Nested* interventions force a variable to take the *natural* state it would have attained under a different, hypothetical intervention (e.g., $Y_{W_{x_0}}$). A *counterfactual* query contrasts multiple states, either by applying conflicting interventions to the same variable or by comparing an observed state with an intervened state. As these queries involve hypothetical scenarios, they cannot be evaluated directly from $P(\boldsymbol{V})$. However, in some cases, the

[1] Online software available at fairmindcausal.streamlit.app, code and examples at github.com/Erhtric/fairmind-causal-fairness-analysis.

independence relations in $\mathscr{G}$ allow us to reduce a counterfactual query to a function of purely observational queries; such queries are termed as *identifiable* [13].

Causal Effects. For the SFM, consider a binary protected variable with $\mathscr{X} := \{x_0, x_1\}$ and a target state $y \in \mathscr{Y}$. We are typically interested in the following queries, for which we provide the definition and also identification formulae:

$$\mathrm{TV}_{x_0,x_1}(y) \coloneqq P(y \mid x_1) - P(y \mid x_0), \tag{1}$$

$$\mathrm{TE}_{x_0,x_1}(y) \coloneqq P(Y_{x_1} = y) - P(Y_{x_0} = y) = \sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)]\, P(z)\,, \tag{2}$$

$$\mathrm{DE}_{x_0,x_1}(y) \coloneqq P(Y_{x_1,W_{x_0}} = y) - P(Y_{x_0,W_{x_0}} = y) = \sum_{z,w} \big[P(y \mid x_1, z, w) - P(y \mid x_0, z, w)\big] P(w \mid x_0, z) P(z)\,, \tag{3}$$

$$\mathrm{IE}_{x_0,x_1}(y) \coloneqq P(Y_{x_0,W_{x_1}} = y) - P(Y_{x_0,W_{x_0}} = y) = \sum_{z,w} P(y \mid x_0, z, w) \big[P(w \mid x_1, z) - P(w \mid x_0, z)\big] P(z)\,, \tag{4}$$

$$\mathrm{SE}_{x_0,x_1}(y) \coloneqq \mathrm{TV}_{x_0,x_1}(y) - \mathrm{TE}_{x_0,x_1}(y)\,. \tag{5}$$

The values of the above causal AF metrics computed by the corresponding identification formulae are depicted as a Sankey diagram in Fig. 2. Statistical disparities in credit allocation with respect to age are shown by the *total variation* ($\mathrm{TV}_{x_0,x_1}(Y=1)$). Following Eq. (5), TV decomposes into the *total effect* ($\mathrm{TE}_{x_0,x_1}(Y=1)$) and the *spurious effect* ($\mathrm{SE}_{x_0,x_1}(Y=1)$). The diagram shows that the disparity primarily arises from causal factors. The low magnitude of SE indicates that the algorithm successfully isolates property values (Z) when determining credit limits. Because causal influence flows both directly and indirectly, we further decompose the causal component: $\mathrm{TE}_{x_0,x_1}(Y=1) = \mathrm{DE}_{x_0,x_1}(Y=1) - \mathrm{IE}_{x_1,x_0}(Y=1)$. The *direct effect* (DE) isolates the causal influence along direct pathways, by measuring how Y changes when setting $X = x_1$ while keeping mediators in their natural state x_0. Similarly, the *indirect effect* (IE) captures the influence mediated by credit score (W). In this context, the system preferentially weights credit scoring while simultaneously allocating greater resources, i.e., higher credit card limit, to older applicants; this being unacceptable from an AF perspective.

In an extended version of this paper, we generalised this framework to general settings involving multiple mediators and confounders, providing exact decomposition formulae for the specific DE and IE contributions of each variable following [14]. Furthermore, we adapt the identification formulae for non-binary targets by substituting probabilities with expectations [3]. For non-binary features, we evaluate descriptors by averaging effects across protected and non-protected partitions, introducing an optional stepwise decomposition for ordinal features to capture effect evolution along the scale.

3 Fairness Analysis and LLM Reports

The above causal descriptors provide a rich AF characterization of the data. However, deriving an intuitive interpretation from these quantities remains challenging for non-specialists, particularly when confounding and mediated effects operate in opposite directions. Visualisations such as the diagram in Fig. 2 can be appealing, but such diagrams can be constructed only when all effects contribute positively; otherwise, more sophisticated—yet potentially less interpretable—representations are required. To address this limitation, we leverage LLMs.

Overall, our pipeline receives a dataset together with an SFM specification as input. Then a Bayesian net is trained from the data to compactly represent the distribution $P(\boldsymbol{V})$ [7]. This allows the computation of observational queries required by the computation of the AF descriptors through the identification formulae in Eqs. (1)–(5).

Finally, the outputs computed by FAIRMIND are passed to an LLM-based reporting module through a dual-prompt architecture. The system prompt specifies the model's role as a statistician and causal-fairness expert, together with the required report structure, formatting and interpretation rules. The user prompt contains only a structured JSON representation of the precomputed fairness results. All statistical calculations remain deterministic and external to the LLM, whose role is limited to interpreting the results and producing a coherent natural-language report. For the following example, we use OpenAI's *gpt-5-mini* [12] with high reasoning effort, providing a balance between report quality, overall token usage and response timings. Fig. 3 depicts the report generated by the FAIRMIND pipeline for the toy example.

FAIRMIND Report
Older applicants ($X = 1$) have a higher probability of receiving a high credit limit ($Y = 1$) than young applicants ($X = 0$). Overall, the observed difference is $TV = 0.4239$ and the causal total effect is $TE = 0.3710$ in the $X = 0 \to X = 1$ direction. The disparity is mainly driven by the indirect (mediated) component associated with W (credit history): the reported reversed-contrast indirect effect is -0.2283, corresponding to an indirect contribution of about $+0.2283$ in the $x_0 \to x_1$ orientation, which is larger than the direct effect ($DE = 0.1427$). Spurious/confounding contribution ($SE = 0.0529$) is relatively small. This suggests differences in credit history account for the largest share of the causal disparity in assigned credit limits by age.

Fig. 3. LLM-powered FAIRMIND report for the toy example.

The reliability of reports generated directly by the LLM is evaluated by verifying that their conclusions preserve the signs, relative magnitudes and causal interpretation of the precomputed effects. Crucially, we restrict the use of LLMs to interpretation rather than calculation. The exact computation of the causal metrics allows us to define a robust pipeline. Furthermore, this helps preserving data privacy and reducing token consumption, as the prompt receives only aggregated causal summaries rather than sensitive raw data.

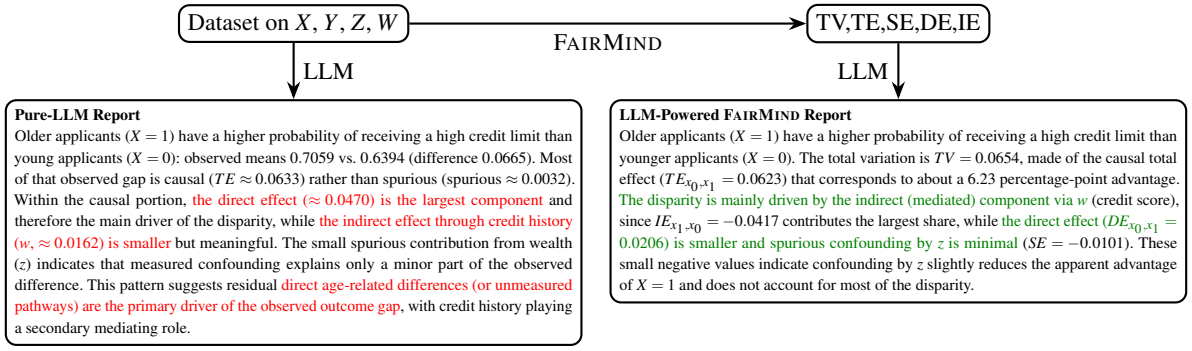


Fig. 4. Report directly generated from the data by LLM vs. the one preprocessed by FAIRMIND. Valid statements are highlighted in green, while invalid statements are highlighted in red.

When variables are binary, as in our initial toy example, both FAIRMIND and the pure LLM approaches yield comparable numerical and text outputs. However, as the state space expands (i.e., $|\mathscr{W}| > 2, |\mathscr{Z}| > 2$), such approach fails to reproduce effects' sizes reliably and eventually arriving to erroneous conclusions. Fig. 4 highlights this

divergence in a more complex banking scenario. We further extend this comparison across model configurations and increasing structural complexity (see Tab. 2), evaluating both pipelines on reliability—correctness of the computed effects, sign, causal interpretation, and overall conclusion validity—and efficiency, measured by token consumption and execution time. Results show that FAIRMIND is sound across tested configurations; the pure LLM pipeline shows inconsistent results across experimental configurations: wrong signs, skewed magnitudes or illogical conclusions. This holds even for lower-tier models, indicating that soundness can be recovered without resorting to costlier reasoning configurations. Within the experimental setup, FAIRMIND reduces token usage by 3–8× and runtime by 3–10× relative to pure LLM, with the gap apparently widening as structural complexity increases.

Table 2. Pure LLM vs. FAIRMIND. **LS**: logically sound, i.e. the report delivers the expected conclusions from the total variation decomposition (Eqs. (1)–(5)). **NS**: numerically sound, i.e. agreement in sign and closeness in magnitude with the symbolic module's computed values. The *categorical mediators* experimental setup consider two mediators, with either four or five states, and two binary confounders. The OPENAI 5.4 and 5.4-nano frontier models are used for experimentation.

Setup	**Config** (LLM model/effort)	**LS**	**NS**	**Tokens I/O/Thinking (kilotokens)**	**Time (s)**
SFM (Fig. 1)	Pure LLM (5.4-nano/low)	✓	✗	7.2/3.4/12.6	63
	Pure LLM (5.4-nano/high)	✓	✗	19.9/19.0/16.8	201
	Pure LLM (5.4/low)	✓	✗	8.3/5.1/13.4	67
	Pure LLM (5.4/high)	✓	✗	8.1/11.2/9.5	141
	FAIRMIND (5.4-nano/low)	✓	✓	2.2/2.0/0.3	17
	FAIRMIND (5.4-nano/high)	✓	✓	2.2/3.7/1.6	32
	FAIRMIND (5.4/low)	✓	✓	2.2/1.8/0.5	13
	FAIRMIND (5.4/high)	✓	✓	2.2/7.4/6.0	57
Categorical mediators	Pure LLM (5.4-nano/low)	✗	✗	12.5/10.5/7.9	140
	Pure LLM (5.4-nano/high)	✓	✗	38.9/37.9/34.9	389
	Pure LLM (5.4/low)	✗	✗	16.0/10.8/8.8	134
	Pure LLM (5.4/high)	✓	✗	41.5/35.5/33.7	415
	FAIRMIND (5.4/low)	✓	✓	2.6/2.1/0.4	19
	FAIRMIND (5.4/high)	✓	✓	2.6/4.7/2.5	37
	FAIRMIND (5.4-nano/low)	✓	✓	2.6/1.9/0.4	24
	FAIRMIND (5.4-nano/high)	✓	✓	2.6/8.9/7.5	74

4 Conclusions

This work introduces FAIRMIND, a robust pipeline that operationalises Plečko and Bareinboim's identifiability results on the SFM [14] to compute exact causal AF metrics and translates them into accessible textual reports using LLMs. We successfully argue that this work promotes the adoption of causal fairness analyses in AutoML settings [1], by requiring minimal user input. To assess robustness of the report generation to varying data-generating settings and increasing model capacity—a current limitation of this short manuscript—future work will include a broader evaluation of the reporting module across different LLMs, repeated runs and multiple datasets. Possible extensions will also include the incorporation of structural learning and causal discovery to automatically identify mediators and confounders from observational data, as well as the possibility to consider data coming from interventional studies. Additionally, we aim to relax the structural assumptions to address partial identifiability and to extend the tool with commonly used automated mitigation strategies for the identified causal biases.

References

[1] Rafael Barbudo, Sebastián Ventura, and José Raúl Romero. 2023. Eight years of AutoML: categorisation, review and trends. *Knowledge and Information Systems* 65, 12 (2023), 5097–5149.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[3] Alessia Berarducci, Eric Rossetto, Alessandro Antonucci, and Marco Zaffalon. 2026. Automatic Causal Fairness Analysis with LLM-Generated Reporting. arXiv:2604.27011 [cs.LG] https://arxiv.org/abs/2604.27011

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[5] Silvia Chiappa and William S. Isaac. 2019. A Causal Bayesian Networks Viewpoint on Fairness. In *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data*, Eleni Kosta, Jo Pierson, Daniel Slamanig, Simone Fischer-Hübner, and Stephan Krenn (Eds.). Vol. 547. Springer International Publishing, 3–20.

[6] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (April 2020), 82–89.

[7] Daphne Koller and Nir Friedman. 2010. *Probabilistic Graphical Models: Principles and Techniques* (nachdr. ed.).

[8] Kriti Mahajan. 2025. CausalFairnessInAction: An Open Source Python Library for Causal Fairness Analysis. In *Proceedings of the NeurIPS 2025 Workshop on Causal Fairness*.

[9] David Martens, James Hinns, Camille Dams, Mark Vergouwen, and Theodoros Evgeniou. 2025. Tell Me a Story! Narrative-Driven XAI with Large Language Models. *Decision Support Systems* 191 (2025), 114402.

[10] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163.

[11] Hamed Nilforoshan, Johann D. Gaebler, Ravi Shroff, and Sharad Goel. 2022. Causal Conceptions of Fairness and Their Consequences. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 16848–16887.

[12] OpenAI. 2026. OpenAI API: Models. https://developers.openai.com/api/docs/guides/reasoning.

[13] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[14] Drago Plečko and Elias Bareinboim. 2024. Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning. *Foundations and Trends in Machine Learning* 17, 3 (2024), 304–589.

[15] Alexander Williams Tolbert. 2025. Why Causal Inference Is Necessary for Algorithmic Fairness. *Synthese* 206, 3 (2025), 162.

[16] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in Decision-Making — The Causal Explanation Formula. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (2018).