AI Explainability and Fairness with Guarantees

Jiří Němeček
contact@nemecekjiri.cz
Czech Technical University
Prague, Czechia

Abstract

AI Fairness and Explainability are two integral parts of Trustworthy AI. However, most definitions of fairness do not consider intersections of protected groups. The ones that do are inefficient, too restrictive, or without guarantees to uncover bias. Similarly, many explainability methods are not guaranteed to return explanations faithful to the model or sacrifice quality to achieve it. We achieve guarantees by utilizing Mixed-Integer Optimization (MIO) in each domain. We introduced MSD, a novel efficient distance measure to guarantee intersectional fairness, and LiCE, a method to obtain the closest counterfactual explanations with sufficient plausibility guaranteed to be faithful to the model. Each of the works suggests future directions. In intersectional fairness, we discuss promising preliminary results in finding a fair classifier. In LiCE, we aim to generalize likelihood estimation to optimization under uncertainty.

CCS Concepts

• **Mathematics of computing** → *Integer programming*.

Keywords

XAI, AI Explainability, Intersectional Fairness, Mixed-Integer Optimization

ACM Reference Format:
Jiří Němeček. 2025. AI Explainability and Fairness with Guarantees. In *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (KDD '25).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Artificial Intelligence (AI) has, in recent decades, grown into an essential technology in multiple domains [13, 28] from finance [4] to education [36]. It became a practical and helpful tool, but at the same time, we have seen multiple, more or less critical, failures of AI systems, often manifesting through discrimination against underrepresented groups. Problems range from misgendering people based on having "Dr" before their name [32], through Amazon's biased hiring tool [8] and Dutch unfair social security fraud detector [1] to the famous COMPAS system skewed in evaluating risk of recidivism [16]. These problems of unequal treatment are tackled by methods of *AI fairness*. AI fairness methods mitigate bias in datasets (or AI model outputs) by equalizing the representation of protected groups in the data (or outputs). There is a wide array of definitions of fairness and even more methods that perform some form of bias mitigation.

In training AI models, fairness methods are usually a form of constrained optimization, minimizing classification error while keeping the bias below some acceptable threshold. On a more individual level, when a user is faced with some decision and suspects bias, they might (have a right to [10, 27]) request an explanation why that decision was reached by the model, to verify that it wasn't influenced dominantly by some protected attributes[1] (e.g., race or sex). AI explainability (XAI) focuses on understanding the predictions of AI models that are difficult to understand directly. In XAI, understanding is achieved through (a series of) *explanations* which abstract away the model's inner workings. Explanations are known to improve user adoption [7, 9] and satisfaction [30], but they are also an essential part of the knowledge discovery pipeline [34]. Many popular explanation methods suffer from various instabilities [18] or limited faithfulness to the underlying model [14, 21]. Counterfactual Explanations (CEs, [33]), which offer a change of input to achieve a desired output, can be stable and are faithful to the model. Still, they are not without issues [31], e.g., offering a single explanation limits the users' autonomy [2].

AI fairness and explainability are sometimes joined under the umbrella term Trustworthy AI [20].

Research Questions. We focus specifically on methods providing guarantees. In either domain, guaranteed quality of an explanation or statistically supported evaluation of bias might not only be desired but also required by regulation or some standard.

To achieve provable guarantees, we utilize *Mixed-Integer Optimization* (MIO, sometimes referred to as Mixed-Integer *Programming*), a method of finding globally optimal solutions [35]. MIO is a form of mathematical optimization where some variables are real-valued, while some can only take integer values. We additionally assume that all formulations have linear constraints, which helps the scalability of the method while still being capable of formulating NP-hard problems [29]. Despite its hardness, the performance of dedicated solvers has advanced rapidly, showing that previously "unsolvable" problems can now be solved in seconds [19].

To effectively utilize MIO, one must consider the (generally) exponential runtime with respect to the size of the formulation (number of constraints and variables). This means that MIO is usually unsuitable for training big ML models since it would require simultaneously representing the training data and the ML model. Thus, we tackle problems where we represent either just the model (to explain it) or just the data (to evaluate a bias measure).

[1] Note that an AI model can be biased even if it doesn't use the protected attributes directly, through correlated features. Checking that a prediction is not made solely on protected attributes is just a necessary condition for fairness.

KDD '25, Toronto, ON, Canada

ACM ISBN 978-1-4503-XXXX-X/2025/08
https://doi.org/XXXXXXX.XXXXXXX

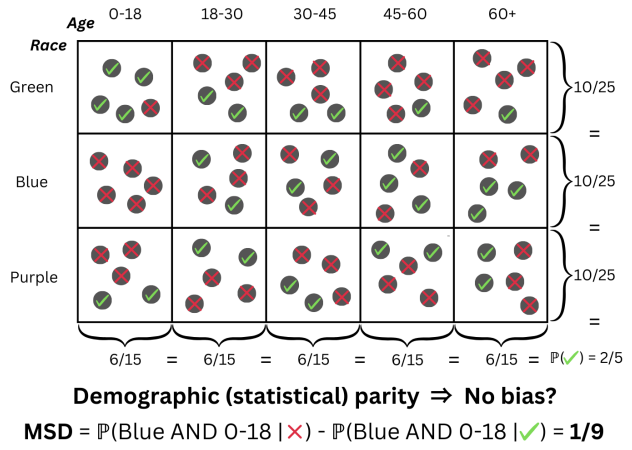

Figure 1: An illustrative example of intersectional bias. Statistical parity is achieved, but the algorithm rejects the entire subgroup of the youngest blue persons. MSD is capable of uncovering this bias and returns the subgroup description.

2 MSD: Maximum Subgroup Discrepancy

In the domain of AI fairness, we consider Intersectional fairness [3] (sometimes called subgroup fairness [22]) which tackles the problem of intersectional bias [6], that is, bias on intersections of marginal groups (e.g., immigrant women above the age of 60, instead of just women or immigrants separately). It has been shown to be undetected by most basic marginal bias metrics like statistical parity or equal opportunity (see Figure 1). Therefore, this form of bias is not mitigated by training methods utilizing those metrics.

Methods of Intersectional fairness exist, but they are often prone to outliers, because they do not consider the number of samples in a given subgroup and rely on enumerating the exponential number of subgroups [5, 11]. Multidimensional subset scanning methods probabilistically sift through all subgroups to find anomalies, which can also be utilized for predictive bias [37], though without a guarantee to find the most disadvantaged group. Finally, Kearns et al. [17] propose SPSF, an effective metric for Intersectional fairness, but consider subgroups to be the output of a linear classifier, which is too restrictive and difficult to interpret.

We propose a novel measure of intersectional bias called Maximum Subgroup Discrepancy (MSD, [23]). It has low sample complexity, efficient computation, and interpretable outputs. Formulating the problem as measuring the distance between two distributions, we show that conventional distance measures have exponential sample complexity. In contrast, our method, built on learning a formula in disjunctive normal form, has linear sample complexity. We show to be able to learn it effectively by using MIO. Finally, we consider subgroups as naturally defined, i.e., intersections (conjunctions) of protected attributes, leaving the user with a description of the most disadvantaged subgroup as an output.

Preliminary results suggest that we are able to optimize small, interpretable models (e.g., linear models) within MIO under the constraint on MSD. We can also formulate the SPSF [17] measure, but on the more interpretable and less restrictive definition of subgroup intersections. We find more accurate predictors than the original SPSF while keeping the fairness violation on the subgroups below a given threshold.

3 LiCE: Likely Counterfactual Explanations

In XAI, we tackle the task of generating plausible Counterfactual Explanations (CEs, [33]). Counterfactual explanations propose how the input would have to change for the output to change as desired. For example, say a customer at a bank is denied a loan. They might get a counterfactual explanation such as: "If your income were 10,000$ higher, your loan would be approved." CEs are well-understandable and can be actionable [31]. Their actionability is limited by some attributes not being subject to change (e.g., race) and by the plausibility of the counterfactual. Plausibility is vaguely defined as the counterfactual not being an outlier in the data distribution [12]. In our example, increasing the income might be feasible, but it might require further education or relocation for better job opportunities. However, the CE does not account for that, and it might be the case that the relocation would also increase the spending on rent, leading to the 10 thousand increase not being enough for the model to reach the desired loan approval anymore. Forcing the counterfactual to have a high likelihood of being sampled from the data distribution can help account for that by making the counterfactual belong to a dense region of the training data.

Our method, LiCE [24], approximates inference of a trained Sum-Product Network (SPN, [26]), a tractable probabilistic model, within MIO to estimate the CE likelihood. We then use MIO to obtain provably closest (most similar to the original input) CEs that are *plausible* by constraining the likelihood above a certain threshold. Alternatively, one can minimize a combination of distance and negative log-likelihood. We show that LiCE generates counterfactuals that are closer and more plausible than counterfactuals generated by other methods that consider plausibility. Moreover, we provide a valid solution in all cases when possible. This is not the case with all other methods, as they sometimes rely on the quality of other trained models [e.g., 25] or hyperparameters that trade-off solution quality for "recall" [e.g., 15].

We consider LiCE a comprehensive method. It includes the enforcement of user-defined constraints, such as causal effects. Further work will thus diverge from the use case of counterfactual explanations and focus on the MIO formulation of a trained SPN, which might prove useful in general optimization under uncertainty.

Conlcusion

We showed that utilizing MIO to obtain solutions with guarantees can be applied to both AI fairness and explainability. In LiCE, we improve over state-of-the-art methods for plausible counterfactuals, using MIO to guarantee that we find the closest CE with a high enough likelihood. With MSD, we propose a method of intersectional fairness, which works with naturally interpretable subgroups, has linear sample complexity with respect to the protected dimension, and can be computed efficiently and optimally with MIO.

Follow-up work, extending the MSD detection of bias to bias *mitigation* already shows promise. We use MIO to find more accurate fair models compared to the existing baseline [17]. This could be a crucial step towards *intersectionally* fair AI models.

References

[1] Amnesty International. 2021. Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. https://www.amnesty.org/en/documents/eur35/4686/2021/en/

[2] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 80–89. doi:10.1145/3351095.3372830

[3] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html ISSN: 2640-3498.

[4] Sean Shun Cao, Wei Jiang, Lijun (Gillian) Lei, and Qing (Clara) Zhou. 2024. Applied AI for finance and accounting: Alternative data and opportunities. *Pacific-Basin Finance Journal* 84 (April 2024), 102307. doi:10.1016/j.pacfin.2024.102307

[5] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2024. Fairness Improvement with Multiple Protected Attributes: How Far Are We?. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. ACM, Lisbon Portugal, 1–13. doi:10.1145/3597503.3639083

[6] Kimberle Crenshaw. 1998. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics. In *Feminism And Politics: Oxford Readings In Feminism*, Anne Phillips (Ed.). Oxford University Press, 0. doi:10.1093/oso/9780198782063.003.0016

[7] Mohammad Dalvi-Esfahani, Mehdi Mosharaf-Dehkordi, Lam Wai Leong, T. Ramayah, and Abdulkarim M. Jamal Kanaan-Jebna. 2023. Exploring the drivers of XAI-enhanced clinical decision support systems adoption: Insights from a stimulus-organism-response perspective. *Technological Forecasting and Social Change* 195 (Oct. 2023), 122768. doi:10.1016/j.techfore.2023.122768

[8] Jeffrey Dastin. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (Oct. 2018). https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

[9] Ashley Deeks. 2019. The Judicial Demand for Explainable Artificial Intelligence. *Columbia Law Review* 119, 7 (2019), 1829–1850. https://www.jstor.org/stable/26810851 Publisher: Columbia Law Review Association, Inc..

[10] Equal Credit Opportunity Act (ECOA) 1974. Equal Credit Opportunity Act (ECOA). https://www.law.cornell.edu/uscode/text/15/chapter-41/subchapter-IV

[11] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An Intersectional Definition of Fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. 1918–1921. doi:10.1109/ICDE48307.2020.00203 ISSN: 2375-026X.

[12] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* (April 2022). doi:10.1007/s10618-022-00831-6

[13] Abid Haleem, Mohd Javaid, Mohd Asim Qadri, Ravi Pratap Singh, and Rajiv Suman. 2022. Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks* 3 (Jan. 2022), 119–132. doi:10.1016/j.ijin.2022.08.005

[14] Xuanxiang Huang and Joao Marques-Silva. 2024. On the failings of Shapley values for explainability. *International Journal of Approximate Reasoning* 171 (Aug. 2024), 109112. doi:10.1016/j.ijar.2023.109112

[15] Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. 2024. Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. In *Proceedings of the 15th Asian Conference on Machine Learning*. PMLR, 582–597. https://proceedings.mlr.press/v222/jiang24a.html ISSN: 2640-3498.

[16] Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2564–2572. https://proceedings.mlr.press/v80/kearns18a.html ISSN: 2640-3498.

[18] Patrick Knab, Sascha Marton, Udo Schlegel, and Christian Bartelt. 2025. Which LIME should I trust? Concepts, challenges, and solutions. In *Proceedings of the XAI 2025 conference*. https://arxiv.org/abs/2503.24365 arXiv: 2503.24365 [cs.LG].

[19] Thorsten Koch, Timo Berthold, Jaap Pedersen, and Charlie Vanaret. 2022. Progress in mathematical programming solvers from 2001 to 2020. *EURO Journal on Computational Optimization* 10 (Jan. 2022), 100031. doi:10.1016/j.ejco.2022.100031

[20] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From Principles to Practices. *Comput. Surveys* 55, 9 (Jan. 2023), 177:1–177:46. doi:10.1145/3555803

[21] Joao Marques-Silva and Xuanxiang Huang. 2024. Explainability Is Not a Game. *Commun. ACM* 67, 7 (July 2024), 66–75. doi:10.1145/3635301

[22] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (July 2021), 115:1–115:35. doi:10.1145/3457607

[23] Jiří Němeček, Mark Kozdoba, Illia Kryvoviaz, Tomáš Pevný, and Jakub Mareček. 2025. Bias Detection via Maximum Subgroup Discrepancy. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD'25)*.

[24] Jiří Němeček, Tomáš Pevný, and Jakub Mareček. 2025. Generating likely counterfactuals using sum-product networks. In *The thirteenth international conference on learning representations*. https://openreview.net/forum?id=rGyi8NNqB0

[25] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 3126–3132. doi:10.1145/3366423.3380087

[26] Hoifung Poon and Pedro Domingos. 2011. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 689–690. doi:10.1109/ICCVW.2011.6130310

[27] Regulation (EU) 2024/1689. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council. of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Place: OJ L, 2024/1689, 12.7.2024.

[28] Mohammed Yousef Shaheen. 2021. Applications of Artificial Intelligence (AI) in healthcare: A review. *ScienceOpen Preprints* (Sept. 2021). doi:10.14293/S2199-1006.1.SOR-.PPVRY8K.v1 Publisher: ScienceOpen.

[29] Hossein Shahrabi Farahani and Jens Lagergren. 2013. Learning Oncogenetic Networks by Reducing to Mixed Integer Linear Programming. *PLoS ONE* 8, 6 (June 2013), e65773. doi:10.1371/journal.pone.0065773

[30] Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. 2022. Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1330–1340. doi:10.1145/3531146.3533189

[31] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. 2024. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *ACM Comput. Surv.* 56, 12 (Oct. 2024), 312:1–312:42. doi:10.1145/3677119

[32] James Vincent. 2020. A service that uses AI to identify gender based on names looks incredibly biased. https://www.theverge.com/2020/7/29/21346310/ai-service-gender-verification-identification-genderify

[33] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal* (2017). doi:10.2139/ssrn.3063289

[34] Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. 2023. Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Information Fusion* 92 (April 2023), 154–176. doi:10.1016/j.inffus.2022.11.013

[35] Laurence A. Wolsey. 2021. *Integer programming* (second edition ed.). Wiley, Hoboken, NJ.

[36] Weiqi Xu and Fan Ouyang. 2022. The application of AI technologies in STEM education: a systematic review from 2011 to 2021. *International Journal of STEM Education* 9, 1 (Sept. 2022), 59. doi:10.1186/s40594-022-00377-5

[37] Zhe Zhang and Daniel B. Neill. 2017. Identifying Significant Predictive Bias in Classifiers. doi:10.48550/arXiv.1611.08292 arXiv:1611.08292 [cs, stat].

Received 15 May 2025