Ensuring Reliable and Transparent Algorithmic Fairness Through Optimal Transport and Uncertainty Quantification

Agathe Fernandes Machado
Université du Québec à Montréal (UQAM), Montreal, Canada
fernandes_machado.agathe@courrier.uqam.ca

Abstract

Machine learning (ML) models are increasingly used in high-stakes decisions, such as insurance pricing and pretrial detention, but often reproduce or amplify biases present in data. To mitigate discrimination, optimal transport (OT) offers a principled way to transform unfair model predictions into fair ones while minimizing performance loss. Moreover, uncertainty-based methods like calibration help assess fairness across sensitive groups, while uncertainty attribution helps identify sources of bias. This research aims to address algorithmic fairness challenges by developing evaluation and mitigation techniques with theoretical guarantees from OT, easily deployable in practice, while integrating fairness into the broader framework of trustworthy AI—enhancing calibration and uncertainty attribution methods to ensure ethical use of ML models by transparency and reliability.

1 Introduction

Although machine learning (ML) models are increasingly used in decision-making areas like pretrial detention, their predictions often reproduce or amplify existing biases from training data, even without intentional design. Several cases of harm to minority groups have been reported in the media, such as facial recognition systems that perform poorly on women with darker skin. In response, many approaches have been proposed to mitigate discrimination, using group fairness metrics (comparing outcomes across sensitive attributes like gender or race) and individual fairness metrics (focusing on a specific individual). Post-processing techniques, which adjust predictions after model training, offer an efficient mitigation strategy without altering the data or model structure.

Of particular importance for measuring and correcting unfairness in model predictions, optimal transport (OT) offers a principled framework to compare score distributions across protected groups and map one to another. Intuitively, achieving fairness in predictions involves transforming unfair scores into fair ones while minimizing the loss of predictive accuracy, with OT offering theoretical guarantees for fairness–performance trade-offs and interpretability via closed-form mappings for continuous univariate distributions.

Ultimately, algorithmic fairness integrates into the broader framework of trustworthy AI, which also emphasizes transparency and reliability in model outputs to ensure its ethical use. Model reliability can be assessed using calibration, which measures how well predictions align with actual outcomes, and is especially useful for fairness when analyzed across sensitive groups. Moreover, uncertainty attribution helps uncover whether sensitive features influence model confidence, revealing potential sources of bias. Therefore, while not limited to fairness, improving uncertainty attribution methods and calibration analysis can further support algorithmic fairness in downstream applications. To address the aforementioned challenges of algorithmic fairness within the broader framework of trustworthy AI, we outline the following research questions:

RQ1: How to develop practical, interpretable tools with theoretical guarantees from OT to ensure fairness across multiple sensitive attributes (MSA)?

RQ2: How can counterfactual fairness be assessed while maintaining feature causal relationships?

RQ3: How can we understand and characterize the gap between achieving calibrated confidence scores from binary classifiers and recovering the true data distribution? How can overall model calibration analysis improve fairness metrics?

RQ4: Can uncertainty attribution (UA) help reduce discrimination in ML models by identifying bias sources?

2 Contribution

2.1 Algorithmic Fairness through Optimal Transport

RQ1: EquiPy, a Python Package for Sequential Fairness with OT Building on the work of [Hu *et al.*, 2024], we developed a Python package, `EquiPy` [Machado *et al.*, 2025b], to address unfairness in ML model predictions regarding MSA, by using Wasserstein barycenters from OT. This package focuses on the Demographic Parity group fairness metric, which aims to ensure similar prediction distributions across sensitive groups. As a post-processing technique, `EquiPy` achieves fairness across MSA by iteratively treating each sensitive attribute, as opposed to enforcing intersectional fairness, allowing the use of closed-form OT solutions. While many packages address algorithmic fairness, our model-agnostic method ensures both strict and approximate fairness across MSA in an interpretable way. It also includes a custom visualization module to help non-technical stakeholders assess and compare various fairness correction strategies by displaying iterative trade-offs between fairness and performance. Finally, the method only requires model predictions and sensitive variables—without access to labels—making it easy to integrate into existing pipelines.

RQ2: Assessing Counterfactual Fairness via OT The concept of "counterfactual fairness" addresses questions such as, "If the protected attributes of an individual had been different, would the prediction have remained the same?", which requires fairness at the individual level rather than at the group level, as in the previous approach. In [Machado *et al.*, 2024b], we link two existing methods for deriving counterfactuals to evaluate counterfactual fairness: adaptations based on a causal graph with quantile preservation [Plečko and Meinshausen, 2020], and multivariate OT [Lara *et al.*, 2024]. We extend "Knothe's rearrangement" and "triangular transport" to probabilistic graphical models and establish the theoretical foundations of a counterfactual approach, called sequential transport, to discuss individual-level fairness. This work led to further exploration of counterfactual calculation for categorical features, as developed in [Machado *et al.*, 2025a].

2.2 RQ3: Enhancing Group Fairness through Calibration

Calibration evaluates the confidence scores produced by probabilistic classifiers. A binary classifier is calibrated if, among all samples assigned a given confidence score, the proportion of positive outcomes equals that score.

Measuring Unfairness through Calibration Errors The widespread use of geospatial data in AI decision-making can perpetuate historical socio-economic biases and exclusionary practices. Moreover, the growing availability of granular spatial data raises ethical issues, as choices in aggregation—whether for scaling or legal compliance—may not be inherently neutral and can hide certain aspects inherent in the data. To address this, we propose a toolkit for detecting biases across geographic areas with varying levels of granularity [Machado *et al.*, 2024c], evaluating fairness by requiring similar calibration errors across sensitive groups. We extend traditional fairness definitions by including spatial sensitive attributes beyond the usual binary ones.

Global Model Calibration for Classifiers While the previous work incorporated calibration into the fairness framework, we aimed to explore global calibration for binary classifiers in order to enhance fairness assessment across sensitive groups in subsequent work. In sensitive fields like healthcare and insurance, calibration is essential since model-predicted scores are often interpreted as event probabilities. Calibrating a classifier improves the alignment between predicted scores and actual outcomes, but a common misconception is treating these scores as true posterior probabilities. Specifically, tree-based classifiers may exhibit low calibration error and high predictive performance, even if their score distributions don't align with true event distributions. In [Machado *et al.*, 2024a], we assess the impact of calibration correction techniques like Platt scaling on the alignment between XGBoost binary classifier scores and true probabilities from simulated data. Results show that while these methods improve calibration, they can increase the divergence between score distribution and true event probabilities.

3 Future Directions

RQ3 Building on our work on global calibration of binary classifiers, we aim to deepen the understanding of calibration as a group fairness metric, focusing on score distribution within minority groups. Limited heterogeneity in these groups may obscure disparities despite low calibration errors.

RQ4 We aim to identify unfairness sources by applying UA techniques, shifting focus from model predictions to their uncertainty. Using game theory such as the Shapley allocation strategy, we plan to analyze the uncertainty in ML predictions to better understand factors contributing to unfair outcomes.

Acknowledgments

I acknowledge OBVIA's funding support for this project.

References

[Hu *et al.*, 2024] François Hu, Philipp Ratz, and Arthur Charpentier. A sequentially fair mechanism for multiple sensitive attributes. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024.

[Lara *et al.*, 2024] Lucas De Lara, Alberto González-Sanz, Nicholas Asher, Laurent Risser, and Jean-Michel Loubes. Transport-based counterfactual models. *Journal of Machine Learning Research*, 25(136):1–59, 2024.

[Machado *et al.*, 2024a] Agathe Fernandes Machado, Arthur Charpentier, Emmanuel Flachaire, Ewen Gallic, and Francois Hu. Post-calibration techniques: Balancing calibration and score distribution alignment. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*, 2024.

[Machado *et al.*, 2024b] Agathe Fernandes Machado, Arthur Charpentier, and Ewen Gallic. Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness. In *The 39th Annual AAAI Conference on Artificial Intelligence*, 2024.

[Machado *et al.*, 2024c] Agathe Fernandes Machado, François Hu, Philipp Ratz, Ewen Gallic, and Arthur Charpentier. Geospatial disparities: A case study on real estate prices in Paris, 2024.

[Machado *et al.*, 2025a] Agathe Fernandes Machado, Arthur Charpentier, and Ewen Gallic. Optimal transport on categorical data for counterfactuals using compositional data and Dirichlet transport, 2025.

[Machado *et al.*, 2025b] Agathe Fernandes Machado, Suzie Grondin, Philipp Ratz, Arthur Charpentier, and François Hu. Equipy: Sequential fairness using optimal transport in Python, 2025.

[Plečko and Meinshausen, 2020] Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020.