

Topic for Bachelor's Thesis

Title: Fairness benefits of Explainable AI

Supervisor: Alexander Schiller

Contact: alexander.schiller1@wu.ac.at

Starting period: immediately

Description:

Methods from Explainable AI serve to provide transparency to predictions made by so-called "black box" machine learning models. Since these methods thus allow the analysis of the decision logic of machine learning models, they could potentially also be used to assess the (un)fairness of the machine learning models, which is the topic of current scientific debate. The goal of this thesis is to a) provide an overview of works exploring the use of Explainable AI for fairness benefits based on a structured literature review and to b) apply a selected Explainable AI method to a classification model on an established dataset, highlighting the gained insights.

Technology: E.g. Python

Literature to start with:

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, 33(1), 26.

Deck, L., Schoeffler, J., De-Arteaga, M., & Kühl, N. (2024). A critical survey on fairness benefits of explainable AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1579-1595).

Papanikou, V., Karidi, D. P., Pitoura, E., Panagiotou, E., & Ntoutsi, E. (2025). Explanations as bias detectors: A critical study of local post-hoc XAI methods for fairness exploration. *arXiv preprint arXiv:2505.00802*.

Please read the [information on our website](#) carefully before starting your thesis!