

Topic for Bachelor's Thesis



Title: A critical perspective on 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. However, the use of Explainable AI methods can be problematic for a variety of reasons such as explanations not being faithful regarding the actual decision logic of the explained model. Additionally, Explainable AI can provoke undesired side effects (e.g., overreliance on machine learning predictions). The goal of this thesis is to a) provide an overview of critical assessments of Explainable AI based on a structured literature review and to b) derive design principles to mitigate risks of employing Explainable AI.

Literature to start with:

Bauer, K., Von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, 34(4), 1582-1602.

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.

Martens, D., Shmueli, G., Evgeniou, T., Bauer, K., Janiesch, C., Feuerriegel, S., ... & Provost, F. (2026). Beware of "explanations" of AI. *Business & Information Systems Engineering*, in press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.

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