Contribution ID: **19** Type: **not specified**

Explainability of Neural Networks in STEM fields

Friday, 20 February 2026 11:30 (1 hour)

Machine learning (ML) models are increasingly deployed in high-stakes environments, e.g., in the health domain, where ethical and legal considerations require models to be interpretable. Despite substantial progress in interpretable ML (IML), several key challenges remain. These include distinguishing between interpreting the model and using the model to interpret the data-generating process, dealing with heterogeneous data, and providing meaningful recourse recommendations. I argue that we can address these challenges by adopting a causal perspective on IML. In particular, I demonstrate how the causal structures of the model and the data-generating process jointly affect model interpretation techniques. I then provide an overview of our recent work on leveraging causal techniques to improve model interpretation and render recourse recommendations more meaningful.

Presenter: GROSSE-WENTRUP, Moritz