26/06/24

9:00 – 9:30: Introduction
9:30 – 10:00: Simon Hanrath: "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." pdf
10:00 – 10:30: Raul Grau: "Concept bottleneck models." pdf

10:30 – 10:45: coffee break

10:45 – 11:15: Balázs Szabados: "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." pdf
11:15 – 11:45: Yoav Rabinovich: "Axiomatic attribution for deep networks." pdf
---
11:45 – 12.45: lunch break

12:45 – 13:15: Bora Kargi: "Understanding global feature contributions with additive importance measures." pdf
13:15 – 13:45: Manuel Arns: "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." pdf
---
13:45 – 14:00: coffee break

14:00 – 14:30 Sam Liang: "A benchmark for interpretability methods in deep neural networks." pdf
14.30 – 15:00: Paulina Stark: "Algorithmic recourse: from counterfactual explanations to interventions." pdf

03/07/24

9:00 – 9:30: Fryderyk Mantiuk: "Progress measures for grokking via mechanistic interpretability." pdf
9:30 – 10:00: Ali Zhunis: "Causal abstractions of neural networks." pdf
---
10:00 – 10:15: coffee break

10:15 – 10:45: Clara Grotehans: "Framework for evaluating faithfulness of local explanations." pdf
10:45 – 11:15: Andrey Khomutov: "Feature relevance quantification in explainable AI: A causal problem." pdf
---
11:15 – 11:30: coffee break

11:30 – 12:00: Marco Wolfer: "Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance." pdf
12:00 – 12:30: Aakarsh Nair: "Full-gradient representation for neural network visualization." pdf
---
12:30 – 13:30: lunch break

13:30 – 14:00: Swagatam Haldar: "The disagreement problem in ele machine learning: A practitioner's perspective." pdf
14:00 – 14:30: Darina Koishigarina: "Explanations can be manipulated and geometry is to blame." pdf
---
14:30 – 14:45: coffee break

14:45 – 15:15: Michael Maier: "Uncovering expression signatures of synergistic drug response using an ensemble of explainable AI models." pdf

15:15 – 15:45: Adhiraj Ghosh: "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection." pdf