

Human-Centric Machine Learning Seminar (Winter 2025)

Topics and papers

1. Interpretable Machine Learning

- i. [Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations](#)
- ii. [Rethinking Interpretability in the Era of Large Language Models](#)
- iii. [Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students?](#)
- iv. [A Diagnostic Study of Explainability Techniques for Text Classification](#)

2. Human Machine Collaboration

- i. [Human Expertise in Algorithmic Prediction](#)
- ii. [Integrating Expert Judgment and Algorithmic Decision Making: An Indistinguishability Framework](#)
- iii. [Human Uncertainty in Concept-Based AI Systems](#)
- iv. [Investigating Agency of LLMs in Human-AI Collaboration Tasks](#)

3. Bias and Fairness

- i. [The Value of Prediction in Identifying the Worst-Off](#)
- ii. [On the Actionability of Outcome Prediction](#)
- iii. [Robust ML Auditing using Prior Knowledge](#)
- iv. [Fairness in Large Language Models: A Taxonomic Survey](#)

4. Algorithmic Auditing in Machine Learning

- i. [Unsolved Problems in ML Safety](#)
- ii. [Red Teaming Language Models with Language Models](#)
- iii. [Are You Getting What You Pay For? Auditing Model Substitution In LLM APIs](#)
- iv. [Is Your LLM Overcharging You? Tokenization, Transparency, and Incentives](#)

5. Uncertainty Quantification in Machine Learning

- i. [Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation](#)
- ii. [Large Language Models Must Be Taught to Know What They Don't Know](#)
- iii. [Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs](#)
- iv. [Uncertainty-Based Abstention in LLMs Improves Safety and Reduces Hallucinations](#)