Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. Dylan Z Slack, Sophie Hilgard, Sameer Singh, Himabindu Lakkaraju. NeurIPS 2021

Learning Cost-Effective and Interpretable Treatment Regimes. Himabindu Lakkaraju, Cynthia Rudin Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54:166-175, 2017.

"How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. AIES 2021.

Robust and Stable Black Box Explanations. Himabindu Lakkaraju, Nino Arsov, Osbert Bastani Proceedings of the 37th International Conference on Machine Learning, PMLR 119:5628-5638, 2020.

Counterfactual Explanations Can Be Manipulated. Dylan Z Slack, Sophie Hilgard, Himabindu Lakkaraju, Sameer Singh. NeurIPS 2021

Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. 26th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, 318–328.

Evaluating the Interpretability of Generative Models by Interactive Reconstruction. Andrew Slavin Ross, Nina Chen, Elisa Zhao Hang, Elena L. Glassman, Finale Doshi-Velez. Arxiv report

Enforcing Interpretability and its Statistical Impacts: Trade-offs between Accuracy and Interpretability. Gintare Karolina Dziugaite, Shai Ben-David, Daniel M. Roy.

Incorporating Interpretable Output Constraints in Bayesian Neural Networks. Wanqian Yang, Lars Lorch, Moritz A. Graule, Himabindu Lakkaraju, Finale Doshi-Velez. NeurIPS 2020

A Causal Framework for Explaining the Predictions of Black-Box Sequence-to-Sequence Models David Alvarez-Melis, Tommi S. Jaakkola EMNLP'17: Empirical Methods in Natural Language Processing. 2017.

**Towards Robust Interpretability with Self-Explaining Neural Networks**

**David Alvarez-Melis**, Tommi S. Jaakkola

*NeurIPS'18: Neural Information Processing Systems. 2018.*

On the Robustness of Interpretability Methods. David Alvarez-Melis, Tommi S. Jaakkola.

Human-in-the-Loop Interpretability Prior Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. NeurIPS 2018

Human Decisions and Machine Predictions. Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan. Quarterly journal of economics.

Interpretable Decision Sets: A Joint Framework for Description and Prediction. Himabindu Lakkaraju , Stephen H. Bach , Jure Leskovec. KDD 2016

How can we fool LIME and SHAP? Adversarial Attacks on Post hoc Explanation Methods. Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, Himabindu Lakkaraju.

"Why Should I Trust You?": Explaining the Predictions of Any Classifier. Marco Tulio Ribeiro , Sameer Singh , Carlos Guestrin. KDD 2016

Anchors: High-precision model-agnostic explanations. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. AAAI 2018

Model-Agnostic Interpretability of Machine Learning. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. ICML 2016 Workshop on Human Interpretability in Machine Learning

Reliable Post hoc Explanations: Modeling Uncertainty in Explainability.

Algorithmic Recourse in the Wild: Understanding the Impact of Data and Model Shifts. Kaivalya Rawal, Ece Kamar, Himabindu Lakkaraju. Arxiv

Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, Himabindu Lakkaraju. Arxiv

An Empirical Study of the Trade-Offs Between Interpretability and Fairness. Jabbari

Et. Al. ICML Workshop2021.