Fairness und Diskriminierungsfreiheit aus Sicht von Ethik und Informatik

Hauptseminar im Sommersemester 2019 – Themen und Einstiegsliteratur

01.) Discrimination in the Philosophical Debate

- Fair Prediction with Disparate Impact: A Study in Recidivism Prediction Instruments. Alexandra Chouldechova. ArXiv 2017.
- Nothing Personal: On Statistical Discrimination. Kasper Lippert-Rasmussen. The Journal of Political Philosophy 15(4) 2007.

02.) Discrimination as a Legal Term

- Discrimination, Artificial Intelligence, and Algorithmic Decision-Making. Frederik Z. Borgesius. Council of Europe 2018.

03.) Racial Profiling

- Racial Profiling. Mathias Risse and Richard Zeckhauser. Philosophy & Public Affairs 32(2) 2004.
- Racial Profiling: A Response to Two Critics. Mathias Risse. Faculty Research Working Paper Series (Harvard University) 2006.
- Why Racial Profiling is Hard to Justify: A Response to Risse and Zeckhauser. Annabelle Lever. Philosophy & Public Affairs 33(1) 2005.
- Racial Profiling and the Political Philosophy of Race. Annabelle Lever. The Oxford Handbook of Philosophy and Race 2016.

04.) Examples and Reasons for Unwanted Algorithmic Discrimination

- Algorithmic Bias in Autonomous Systems. David Danks and Alex J. London. IJCAI 2017.
- Investigating the Impact of Gender on Rank in Resume Search Engines. Le Chen et al. CHI 2018.
- Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. Reuben Binns et al. SocInfo 2017.
- Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. Juhi Kulshrestha et al. CSCW 2017.
- The Frontiers of Fairness in Machine Learning. Alexandra Chouldechova and Aaron Roth. ArXiv 2018.
- A Framework for Understanding Unintended Consequences of Machine Learning. Harini Suresh and John V. Guttag. ArXiv 2019.

05.) Techniques to Discover and Evaluate Algorithmic Discrimination

- Controlling Machine-Learning Algorithms and Their Biases. Tobias Baer and Vishnu Kamalnath. McKinsey Article 2017.
- A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. Till Speicher et al. KDD 2018.
- Fairness Testing: Testing Software for Discrimination. Sainyam Galhotra et al. SIGSOFT 2017.
- Integrating Induction and Deduction for Finding Evidence of Discrimination. Salvatore Ruggieri et al. Artif. Intell. & Law 2010.
- Unbiased Look at Dataset Bias. Antonio Torralba and Alexei A. Efros. CVPR 2011.
- Problem Formulation and Fairness. Samir Passi and Solon Barocas. FAT* 2019.

06.) Formalizing Fairness and Non-Discrimination

- Fairness Definitions Explained. Sahil Verma and Julia Rubin. FairWare 2018.
- An Intersectional Definition of Fairness. James Foulds and Shimei. ArXiv 2018.
- Discriminative but Not Discriminatory: A Comparison of Fairness Definitions Under Different Worldviews. Samuel Yeom and Michael C. Tschantz. ArXiv 2019.
- On Formalizing Fairness in Prediction with Machine Learning. Pratik Gajane and Mykola Pechenizkiy. ArXiv 2018.
- Fairness Risk Measures. Robert C. Williamson and Aditya K. Menon. ArXiv 2019.
- Fairness Through Awareness. Cynthia Dwork et al. ITCS 2012.

07.) Process, Outcome and Counterfactual Fairness

- Counterfactual Fairness. Matt J. Kusner et al. NIPS 2017.
- Fairness in Decision-Making - The Causal Explanation Formula. Junzhe Zhang and Elias Bareinboim. AAAI 2018.
- When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. Chris Russell et al. NIPS 2017.
- Fair Inference on Outcomes. Razieh Nabi and Ilya Shpitser. AAAI 2018.

08.) Trade-Offs and the Cost of Fairness

- On the (Im)Possibility of Fairness. Sorelle A. Friedler et al. ArXiv 2016.
- The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. Sam Corbett-Davies and Sharad Goel. ArXiv 2018.
- Inherent Trade-Offs in the Fair Determination of Risk Scores. Jon Kleinberg et al. ArXiv 16.
- Does Mitigating ML's Disparate Impact Require Disparate Treatment? Zachary C. Lipton et al. ArXiv 2017.
- Algorithmic Decision Making and the Cost of Fairness. Sam Corbett-Davies et al. KDD 2017.

09.) Algorithmic Fairness and Social Equality (Economic Models)

- Fairness for Whom? Critically Reframing Fairness with Nash Welfare Product. Ansh Patel. ArXiv 2018.
- From Parity to Preference-based Notions of Fairness in Classification. Muhammad B. Zafar et al. NIPS 2017.
- Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. Hoda Heidari et al. NIPS 2018.
- From Fair Decision Making To Social Equality. Hussein Mouzannar et al. FAT* 2019.
- A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity. Hoda Heidari et al. FAT* 2018.

10.) Causal Reasoning for Fairness

- Avoiding Discrimination through Causal Reasoning. Niki Kilbertus et al. NIPS 2017.
- Causal Reasoning for Algorithmic Fairness. Joshua R. Loftus et al. ArXiv 2018.
- Causal Modeling-Based Discrimination Discovery and Removal: Criteria, Bounds, and Algorithms. Lu Zhang et al. Trans. Knowl. & Data Eng. 2018.
- Fairwashing: The Risk of Rationalization. Ulrich Aïvodji et al. ArXiv 2019.

11.) Approximate Fair Treatment and Sanitizing Classifiers

- Achieving Fair Treatment in Algorithmic Classification. Andrew Morgan and Rafael Pass. TCC 2018.

12.) Composition of Approximate Fair Treatment

- Fairness under Composition. Cynthia Dwork and Christina Ilvento. ITCS 2019.