Thesis References

# Fairness and Machine Learning

* 1. <https://fairmlbook.org/>
  2. Revisit the “leap of faith” in machine learning and interrogate how institutions make decisions about individuals
     1. Do not compare machine learning models to the subjective judgements of individual humans but to institutional decision-making
  3. Fairness: moral lens through which we examine the decisions made in machine learning
     1. Ex. accepting or rejecting job applicants—how do measured characteristics of an individual lead to different outcomes
     2. Discrimination would be the wrongful consideration on the basis of group membership
     3. Discrimination is a domain specific concept as it applies to opportunities that affect individuals

|  |  |  |
| --- | --- | --- |
| Chapter 2 | Properties that make automated decision making legitimate and a concern |  |
| **Chapter 3** | **Statistical decision theory to formalize fairness criteria (statistical fairness criteria has three different definitions based on different moral intuitions)** | **Criteria of fairness, which definition fits, how to compare fairness** |
| **Chapter 4** | **Normative underpinnings of objections to systematic differences in the treatment of different groups and inequalities in the outcome experienced by these groups** | **What makes discrimination bad** |
| Chapter 5 | Introduction to the formal concepts of causality present in technical and legal scholarship on discrimination |  |
| Chapter 6 | Legal |  |
| **Chapter 7** | **Complexities of testing for discrimination in practice through experiments and audits** | **How to test for discrimination, its shortcomings, considerations** |

* 1. Studying discrimination in decision making has been criticized as a narrow perspective on broader system of injustice
     1. Neglects the powerful structural determinants of discrimination (law, policy, infrastructure, …)
     2. Orients space of intervention towards solutions that reforms existing decision making systems (updates to an algorithm) which prioritizes “tech fixes” which are like band-aids in comparison to structural interventions and alternatives to machine learning

# Bias in Computer Systems—Friedman and Nissenbaum

1. <https://dl.acm.org/doi/pdf/10.1145/230538.230561>
2. Three categories of bias in computer systems
   1. Preexisting: roots in social institutions, practices, and attitudes
   2. Technical: arise from technical constraints or considerations
   3. Emergent: arise in context of use
3. **Freedom from bias means freedom from all 3**

# Algorithmic bias: on the implicit biases of social technology—Johnson

1. <https://link.springer.com/article/10.1007/s11229-020-02696-y>
2. Algorithmic bias: bias inherited from social patterns reflected in an algorithm’s training data without any directed effort by programmers to include such biases
3. Relationship between machine bias and human cognitive bias
4. “the proxy problem”: biases revisit revision because they rely on proxy attributes, seemingly harmless attributes that correlate with socially-sensitive attributes, serving as proxies for these socially-sensitive attributes

# On Statistical Criteria of Algorithmic Fairness—Hedden

* 1. <https://onlinelibrary.wiley.com/doi/full/10.1111/papa.12189>
  2. Famous case study of algorithmic fairness: Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm to predict recidivism by assigning risk score to individuals on basis of questionnaire (not including race) with a higher rate of false positives for blacks than for whites AND higher rate of false negatives for whites than blacks
     1. Found not to be bias by developer because predictions were equally accurate for two groups
     2. Found to be bias since different predictive accuracy and calibration
  3. Fair prediction is impossible and moral dilemmas are inevitable—“there is no such thing as an unbiased program”
  4. **Argues none of the criteria of fairness except calibration are a genuine, necessary condition on the fairness of predictive algorithms**

# Algorithmic bias: Senses, sources, solutions—Fazelpour and Danks

1. <https://compass.onlinelibrary.wiley.com/doi/full/10.1111/phc3.12760>
2. Facilitate the application of philosophical analysis to these contested issues by providing an overview of
   * What algorithmic bias is
   * Why and how it can occur
   * And what can and should be done about it

# What’s Fair about Individual Fairness?—Fleisher

* 1. [What's Fair about Individual Fairness? | Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society](https://dl.acm.org/doi/10.1145/3461702.3462621)
  2. Individual fairness (IF) methods motivated by the principle of similar treatment that gives correct definition of algorithmic fairness (according to some)
  3. Argues that individual fairness cannot serve as a definition of fairness and IF should not be given priority over other fairness methods with four fundamental problems of IF
     1. Insufficient guarantee of fairness (counterexamples)
     2. Risks encoding human bias
     3. Requires prior moral judgments
     4. Incommensurability of relevant moral values

# Algorithmic Fairness and Base Rate Tracking—Eva

1. <https://onlinelibrary.wiley.com/doi/full/10.1111/papa.12211>
2. Introduction—example: mortgage application is denied and suspected to be related to disadvantaged group membership, in order to evaluate the fairness of the algorithm on the on the basis of its input and output you need to use statistical criteria of algorithmic fairness
3. Statistical criteria of algorithmic fairness: purely statistical criteria which specify necessary conditions that must be satisfied by an algorithm’s predictions in order for the algorithm to count as fair
   1. Powerful objects and impossibility results
4. The article will present a novel statistical criterion of algorithmic fairness that is resistant to Hedden’s counterexample and equipped to accurately diagnose unfairness like example

|  |  |
| --- | --- |
| Section 2 | Review 11 of the most influential statistical criteria of algorithmic fairness |
| Section 3 | Evaluate proper formulation and limitations of the calibration within groups criterion |
| Section 4 | Introducing novel base rate tracking criterion |
| Section 5 | Defends base rate tracking |

# Reading schedule:

Fairness and Machine Learning (3, 4, 7) –can you get this on your iPad with notes?

Bias in Computer Systems—Friedman and Nissenbaum

On Statistical Criteria of Algorithmic Fairness—Hedden