Algorithmic Fairness

PHI420/PHI591 - Fall 21

Marcello Di Bello - mdibello@asu.edu

Th 3-5:50 PM in Tempe, COOR ~~3301~~ 4401

Office hours: Thursday after class or by appointment

Topic

Public and private sector entities rely on algorithms to streamline decisions about healthcare, loans, social services, sentencing and more. This course examines the ongoing debate among computer scientists, legal scholars and philosophers about the fairness of algorithmic decision-making. What does it mean for an algorithm to be fair? How does algorithmic fairness relate to other ideas such as anti-discrimination and fair equality of opportunity? Can fairness be measured quantitatively? What other values and goals should inform the design of ethical algorithms, such as minimizing the harm toward disadvantaged minorities or respecting each person's individuality?

Objectives

This course is meant for advanced undergraduates and graduates students. You will become familiar with different conceptions of algorithmic fairness, such as predictive parity, classification parity and counterfactual fairness, as well as explore connections between algorithmic fairness, anti-discrimination law and philosophical concepts such as fair equality of opportunity. You will sharpen your critical thinking skills, in reading and writing, for the analysis of new technologies and the ethical issues they pose. You will read academic papers in different disciplines—computer science, economics, law, sociology and philosophy—and develop an appreciation for differences in scholarship across disciplines.

Course Materials

Readings and other course materials are available on this website. For readings covered by copyright, please check the Canvas page for this course.
Course materials are divided into "essential" and "additional". You are only expected to study the essential ones, but I recommend that you have a look at the additional materials for at least one or two weeks.

Requirements

Participation

Please attend class regularly and participate (**10% of our grade**). This is a "seminar style" course. You are expected to take an active role in the discussions. Please study the assigned materials **before class** and be ready to discuss them.

Assignments

In addition, you should write ten Pass/Fail précis as well as three graded essays or a research paper (**90% of your grade**).

Pass/fail

Every week please write a **one-page précis** of one of the papers assigned for that week. The précis should describe: (a) topic of the paper; (b) main thesis (or main theses, if there are more than one); (c) supporting arguments; (d) objections to these arguments, complications or difficulties that the author considers (if any). Submit your précis each week through Canvas **before the beginning of class**. You should receive a PASS in **at least ten précis**, or else a full letter grade will be subtracted from your final grade (i.e. A would become B; B would become C; etc.)

Essays

There will be three main graded assignments for this course, **5 pages each**.

1. After doing independent research, collect and summarize **two stories**. One should describe how an algorithm has positively impacted people's lives and another how an algorithm has negatively impacted people's lives. The two stories need not be—and often will not be—about the same algorithm, but the same algorithm could have both positive and negative effects.

2. Write **an exploratory essay** that describes very carefully (a) ProPublica's accusation that COMPAS is racially biased and Northpointe's response; (b) assess what further technical, practical and theoretical questions about algorithmic fairness must be addressed in order to settle the debate. Please be precise.

3. Write a **philosophical argumentative essay** that attempts to answer the question "What is Algorithmic Fairness?" In answering this question, I recommend that you compare and contrast the formal notions of algorithmic fairness (such as predictive parity, classification parity, causal fairness, etc.) with either (i) the economic literature on discrimination, (ii) equal protection jurisprudence, (iii) the philosophical literature on fairness and equality. Please always defend the claims you make with careful and reasoned arguments. Check out these guidelines and these other guidelines on how to write a philosophical argumentative essay.

Research paper

If you are a grad student or advanced undergrad with research experience, you may combine the three 5 page essays into one **15-20 page** research paper. The research paper should engage closely with a subset of the course materials including the additional ones. It is neither necessary nor recommended that you use materials outside those already included in the course materials.

Please come talk to me before you start working on the research paper.

Schedule

Wk 1 - 8/19

Introduction

The question of algorithmic fairness is technical (how do algorithms work?), practical (how do algorithms impact people's lives?) and theoretical (what does "fairness" mean?). On the technical side, we will survey key concepts, such as algorithm, machine learning, supervised v. unsupervised learning, regression, classification v. clustering. This survey will be informal. Those who wish to go deeper may consult the additional materials. On the practical side, we will read stories about people whose lives were affected, often adversely, by algorithms. On the theoretical side, we will brainstorm ideas about what we mean by "fairness" and how an algorithm can be unfair. We will explore several definitions of fairness more deeply later in the course.

Key readings and materials

- Class notes Wk 1
- O'Neil, Weapons of Math Destruction - Introduction
- Kearns and Roth, The Ethical Algorithm - Chp. 1
- A Gentle Introduction to Machine Learning
- Eubanks, Automating Inequality - Chp. 4

Additional readings and materials

- Rogers and Girolami, A First Course in Machine Learning, Chp. 1
- Nielsen, Neural Networks and Deep Learning - Chp. 1
- What Is Backpropagation Really Doing?
- Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office

Wk 2 - 08/26

ProPublica v. Northpointe

To examine more closely the interplay between technical, theoretical and practical questions, we will look at a case study, the algorithm COMPAS used in criminal justice to help judges make decisions about bail, preventative detention and sentencing. The website of investigative journalism ProPublica argued that COMPAS is biased against black people. Northpointe (now Equivant), the company that designed COMPAS, rejected the accusation claiming that COMPAS is racially fair. They both provided numbers to back up their claims. So is COMPAS racially biased or not? The crux of the matter here seems to be a disagreement about the meaning of "fairness".

To better understand this debate, you should be familiar with a bif of probability theory and machine learning, in particular, notions such as false positive, false negative, sensitivity, specificity, Area Under the Curve (AUC) and confusion matrix. We will review these notions as needed.

Key readings and materials

- Class notes Wk 2
- Machine Learning Fundamentals:
  - Sensitivity and Specificity
  - ROC and AUC
  - Confusion Matrix
- ProPublica, Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.
- Dieterich, Mendoza and Brennan, COMPAS risk scales: Demonstrating Accuracy Equity and Predictive Parity Performance of the COMPAS Risk Scales in Broward County

Additional readings and materials

- ProPublica, How We Analyzed the COMPAS Recidivism Algorithm
  - Exploring the ProPublica COMPAS data
- Northpointe, Practitioner's Guide to COMPAS Core
  - Sample Risk Assessment COMPAS
- Flores, Bechtel and Lowenkamp. False Positives, False Negatives, and False Analyses: A rejoinder to "Machine Bias"
- Parole Denied: One Man's Fight Against a COMPAS Risk Assessment

Wk 3 - 9/2

(Mis)guided by Data

Let's take a step back and examine the basis of algorithmic decision-making: data. Though incomplete and never truly conclusive, data can guide our decisions, often benefitting people. For example, data have been playing a key role in the response to the COVID-19 Pandemic (see,among others, Abedin et al, Using Data to Inform the COVID-19 Policy Response)

Another example, in the context of criminal justice, is the data-driven Public Safety Assessment (PSA) algorithm, now widely used in several jurisdictions in the United States to help judges make pre-trial decisions. According to a report from New Jersey, pre-trial jail population descrease significantly after PSA was adopted. The ACLU of New Jersey endorsed the use of the PSA algorithm because it has the potential to end the pre-trial system of money and bail that disproportionately harms the poor.

Although data can be helpful to make decisions, they can also be biased, partial or incomplete. Inadequate data can be harmful especially for already disadvantaged minorities. An important problem is the use of "proxies" in the historical data. That is, the algorithm is not trained on the actual outcomes (which often cannot be known), but on proxies of the actual outcome (say 'arrest' is used as proxy for 'crime'). We will see a few examples of biased, partial and incomplete data and their harmful effects on people.

Key readings and materials

- Class notes Wk 3 (if interested, check out SVM with ProPublica COMPAS Data)
- Kearns and Roth, The Ethical Algorithm - Chp. 2 (only pp. 57-64)
- Hardt, How Big Data Is Unfair
- Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns
- Perez, Invisible Women: Data Bias in a World Designed for Men - Chp. 8

Additional readings and materials

- Caliskan, Bryson, and Narayanan, Semantics Derived Automatically from Language Corpora Contain Human-Like Biases
- Hart and Cooke (2013), Another Look at the (Im-)precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments
  - Imrey and Dawid (2015), A Commentary on Statistical Assessmentof Violence Recidivism Risk
- Obermeyer et al (2019), Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations
- Lazer, Kennedy, King, and Vespignani, The Parable of Google flu: Traps in Big Data Analysis
- Buolamwini, Algorithms Aren't Racist. Your Skin Is just too Dark

Wk 4 - 9/9

Feedback Loops

Another problem with data is that decisions based on inadequate data magnify errors on a large scale especially when they are automated. Even unbiased data can produce pernicious feedback loops. We will exmaine examples of feedback loops in algorithms used in education and policing.

Key readings and materials

- Class notes Wk 4 (if interested, check out my note on Simulating Feedback Loops in Predictive Policing)
- O'Neil, Weapons of Math Destruction - Chps. 3, 4 and 5
- Lum and Isaac, To Predict and Serve?

Additional readings and materials

- Ensign, Friedler, Neville, Scheidegger and Venkata, Runaway Feedback Loops in Predictive policing - video
- Harcourt, Against Prediction - Chp. 5
- Brantingham et al (2018), Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Control Trial

First essay due

Wk 5 - 9/16

Meanings of Fairness

We have not yet asked the fundamental questions. What does it mean for algorithmic decisions to be fair? How is fairness different from accuracy? Computer scientists have formulated several different metrics of algorithmic fairness. By one count, there are as many as 21 definitions. Three of the most important definitions out there are called: *demographic parity*, *classification parity* and *predictive parity*. We will examine these definitions and their relationship to accuracy.

Key readings and materials

- Class notes Wk 5
- Kearns and Roth, *The Ethical Algorithm*, Chp. 2
- Corbett-Davies and Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning - video

Additional readings and materials

- Narayanan, 21 Fairness Definitions and Their Politics
- Barocas, Hardt and Narayanan, Fairness in Machine Learning - Chp. 2

Wk 6 - 09/23

Impossibility Theorems

It turns out it is mathematically impossible, under realistic conditions, for an algorithm to satisfy different conceptions of fairness at the same time. These impossibility theorems led scholars to consider trade-offs between different conceptions of fairness or be skeptical toward existing definitions.

Q&A with Deborah Hellman

The first part of the class will consist in a Q&A session with Deborah Hellman, Professor of Law at the University of Virginia School of Law and the author of one of the readings assigned for today. *Please come prepared to ask questions.*

Key readings and materials

- Class notes Wk 6 (also check out Chouldechova's Impossibility: In Pictures)
- Hellman, Measuring Algorithmic Fairness
- Chouldechova, Fair Prediction with Disparate Impact

Additional readings and materials

- Grant, Is It Impossible to Be Fair?
- Hedden, On Statistical Criteria of Algorithmic Fairness - video

- Berk, Heidari, Jabbari, Kearns and Roth, Fairness in Criminal Justice Risk Assessments: The State of the Art
- Kleinberg, Mullainathan and Raghavan, Inherent Trade-offs in the Fair Determination of Risk Scores - video

Wk 7 - 09/30

Fair Risk Assessment

The debate about algorithmic fairness is nothing new. Since the development of statistical methods in insurance and banking, claims of racial and gender discrimination have been advanced against risk assessment tools used to price loans and insurance policies. We will review this history to understand analogies and differences between fair risk assessment in the 20th century and algorithmic fairness in the 21st century. We will also look at how ordinary people make distinctions between fair and unfair data-based risk predictions.

Key readings and materials

- Class notes Wk 7
- Ochigame, The Long History of Algorithmic Fairness
- Kiviat, Which Data Fairly Differentiate?

Second essay due

Wk 8 - 10/7

Structural Unfairness

The definitions of algorithmic fairness we have seen so far describe the extent to which *actual* rates of error are unevenly (and thus unfairly) allocated across different socially relevant groups. Scholars in computer science have also used more sophisticated definitions to formulate a *counterfactual* or *causal* definition of algorithmic fairness. These definitions attempt to capture the structural dimension of racial discrimination (as in the expression "structural racism"). Other scholars in law, sociology and philosophy have argued that these attempt are unsatisfactory.

Q&A with Lily Hu

The first part of the class will consist in a Q&A session with Lily Hu, a philosopher at Harvard interested in algorithmic bias and how the social sciences measure the causal effect of race. She is the author of some of the readings assigned for today. *Please come prepared to ask questions.*

Key readings and materials

- Class notes Wk 8
- Hu, What is 'Race' in Algorithmic Discrimination on the Basis of Race?
- Hu, Disparate Causes, Part I and Part II
- Kohler-Hausmann, Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination

Additional readings and materials

- Barocas, Hardt and Narayanan, Fairness and Machine Learning - Chp 5: Causality
- Kusner, Loftus, Russell and Silva, Counterfactual Fairness
- Chiappa and Gillam, Path-specific Counterfactual Fairness

Wk 9 - 10/14

Structural Unfairness (cont'ed)

We will continue our discussion of the structural approach to algorithmic fairness.

Q&A with Annette Zimmermann

The first part of the class will consist in a Q&A session with Annette Zimmermann, a political philosopher at the University of York working on the ethics of artificial intelligence and machine learning and the author of one of the readings assigned for today. *Please come prepared to ask questions.*

Key readings and materials

- Zimmerman & Lee-Stronach, Proceed with Caution

Additional readings and materials

- Green, The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness

Wk 10 - 10/21

Algorithmic Equal Protection

Algorithimic fairneess is closely conected with antidiscrimination and equal protection jurisprudence, specifically the notions of "disparate treatment" and "disparate impact". Any differential treatment that is due to racial animus or discriminatory intent is clearly illegal. But the Supreme Court has allowed the use of race in specified contexts, for example, in college admissions for the purpose of fostering diversity (see e.g. Fisher v. University of Texas (2016)). Evidence of disparate impact against a protected group is enough to make a prima facie case of discrimination. This applies to sectors such as employment and housing (see e.g. Title VII of the Civil Rights Act of 1964 and Hazelwood School District v. United States (1977)). The criminal justice system, however, seems exempt. An elaborate statistical analysis – showing that death penalty decisions in Georgia disproportionately targeted African Americans, controlling for several variables – was not enough to convince the Court that the system violated equal protection (see McClensky v. Kemp (1987)). It is worth thinking about whether the legal notions of disparate treatment and disparate impact can help to understand what algorithmic fairness consists in.

Key readings and materials

- Barocas and Selbst, Big Data's Disparate Impact

Additional readings and materials

- Any court opinion linked above

Wk 11 - 10/28

The Economics of Discrimination

The academic debate about algorithmic fairness mirrors, in some important ways, the debate among economists about different tests for what counts as discriminatory behaviour. We will learn the basic of the economic approach to discrimination guided by Nobel prize winners Gary Becker and Kenneth Arrow

Key readings and materials

- Becker, The Economic Way of Looking at Life (mostly focus on the sections about discrimination and crime)
- Arrow, What Has Economics to Say About Racial Discrimination?
- Stanford Open Policing, Nationwide Analysis of Traffic Stops and Searches

Additional readings and materials

- Pierson et al, A large-scale Analysis of Racial Disparities in Police Stops Across the United States
- Knowles, Persico and Todd, Racial Bias in Motor Vehicle Searches: Theory and Evidence
- Simoiu, Corbett-Davies and Goel, The Problem of Infra-Marginality

Wk 12 - 11/4

Fairness v. Welfare

We now analize the question of algorithmic fairness from the perspective of utilitarianism. Utilitarianism, roughly, says that the right action is one that maximizes social welfare. If utilitarianism is the correct ethical theory, algorithms should maximize social welfare and this may require to treat people unfairly.

Key readings and materials

- Kaplow and Shavell, Fairness Versus Welfare
- Huq, Racial Equity in Algorithmic Criminal Justice

Additional readings and materials

- Bentham, Introduction to the Principles of Morals and Legislation, Chps. I–II

Wk 13 - 11/11

Equality

Against utilitarianism, John Rawls, an influential political philosopher of the 20th century, argued that society must give everyone equal opportunities to develop their talents and that inequalities in the distribution of goods can be tolerated only if they work to the advantage of the worst-off in society. What would algorithmic fairness look like in light of Rawls's view about equal opportunities and distributive justice? Other influential philosophers, Elizabeth Anderson and Tim Scanlon, have also defended egalitarian views. What would algorithmic fairness look like in light of these egaliatarian views?

Key readings and materials

- Rawls, Justice as Fairness, Part 2, Sections 12-18
- Anderson, What is the Point of Equality?
- Binns, Fairness in Machine Learning: Lessons from Political Philosophy

Additional readings and materials

- Scanlon, Why Does Inequality Matter?
- Kolodny, Why Equality of Treatment and Opportunity Might Matter
- Heidari, Loi, Gummadi and Krause, A Moral Framework for Understanding fair ML Through Economic Models of Equality of Opportunity

Wk 14 - 11/18

Individualized Decisions

Since algorithms often rely on statistical generalizations and correlations, some worry that algorithms disregard the individual circumstances of each person. Do people have a right to be treated as individuals? What would that mean?

Key readings and materials

- Lippert-Rasmussen, We Are All Different

Additional readings and materials

- Eidelson, Treating People as Individuals
- Beeghly, Failing to Treat Persons as Individuals

Wk 15 - 11/25

Thanksgiving, No Class

Wk 16 - 12/2

Q&A with Seth Lazar and Jake Stone

Seth Lazar and Jake Stone from the Australian National University will present their ongoing project 'Machine Learning and the Ethics of Inference'

Third essay or research paper due

Similar courses

- Aaron Fraenkel's Fairness and Algorithmic Decision Making - UC, San Diego
- Solon Barocas, Moritz Hardt, Arvind Narayanan, Fairness in Machine Learning - book
- Moritz Hardt's Fairness in Machine Learning - UC, Berkeley
- Arvind Narayanan's Fairness in Machine Learning - Princeton

Title IX

Title IX is a federal law that provides that no person be excluded on the basis of sex from participation in, be denied benefits of, or be subjected to discrimination under any education program or activity. Both Title IX and university policy make clear that sexual violence and harassment based on sex is prohibited. An individual who believes they have been subjected to sexual violence or harassed on the basis of sex can seek support, including counseling and academic support, from the university. If you or someone you know has been harassed on the basis of sex or sexually assaulted, you can find information and resources at https://sexualviolenceprevention.asu.edu/faqs.

As a mandated reporter, I am obligated to report any information I become aware of regarding alleged acts of sexual discrimination, including sexual violence and dating violence. ASU Counseling Services, https://eoss.asu.edu/counseling is available if you wish to discuss any concerns confidentially and privately. ASU online students may access 360 Life Services, https://goto.asuonline.asu.edu/success/online-resources.html.