Overview

Course:	CS 109A: Data Science 1: Introduction to Data Science
Course Level:	Upper-level undergraduate
Course Description:	"Data Science 1 is the first half of a one-year introduction to data science. The course will focus on the analysis of messy, real life data to perform predictions using statistical and machine learning methods. Material covered will integrate the five key facets of an investigation using data: (1) data collection - data wrangling, cleaning, and sampling to get a suitable data set; (2) data management - accessing data quickly and reliably; (3) exploratory data analysis – generating hypotheses and building intuition; (4) prediction or statistical learning; and (5) communication – summarizing results through visualization, stories, and interpretable summaries. Part one of a two-part series. The curriculum for this course builds throughout the academic year. Students are strongly encouraged to enroll in both the fall and spring course within the same academic year."[1]
Module Topic:	Algorithmic (Un)fairness
Module Author:	Sophie Gibert
Semesters Taught:	Fall 2021
Tags:	Algorithms [CS], machine learning [CS], algorithmic design [CS], prediction [CS], bias [phil], discrimination [phil], fairness [phil], transparency [phil], label [CS], training data [CS], proxy variable [CS], feedback loop [phil], calibration [CS], luck-egalitarianism [phil], model cards [CS]
Module Overview:	This module focuses on the questions of when machine-learning (ML) algorithms are unfair and how we can prevent them from being so. The module begins with an overview of the ways in which unfair bias can enter ML algorithms. The module then focuses on a case study involving a racially biased healthcare risk-prediction algorithm. In small groups, students attempt to articulate the sense in which the healthcare algorithm is unfair. They are then introduced to the promises and pitfalls of a luck-egalitarian answer. To close, the module provides a set of ethical questions that students should ask during the design and examination of ML algorithms and introduces the practical tool of a "Model Card for Model Reporting." In the post-class assignment, students practice generating key ethical questions about an algorithm's design.
Connection to Course Material:	Students in this course learn to analyze real data and perform predictions using statistical and machine learning methods. The module poses ethical questions about how to avoid reproducing unfair biases when employing such methods.
	The topic was chosen because of its direct connection to the technical material covered in the course. The topic is also timely: there is an active and growing research program on algorithmic bias, and the healthcare algorithm studied in the module was exposed as racially biased only in 2019. It and other algorithms like

[1] Link.

		it are still widely used today by both public and private entities.

Goals		
Module Goals:	By the end of the module, students should be able to: 1. Identify multiple stages of development at which bias can enter ML algorithms. 2. Describe a major healthcare risk-prediction algorithm and explain how it discriminates against Black patients. 3. Articulate why the outcome of this algorithm is unfair by appealing to luck-egalitarianism; problematize this answer and offer an alternative. 4. Generate ethical questions to ask during the development and evaluation of a given ML algorithm that help to address unfairness.	
Key Philosophical Questions:	1. How can unfair bias arise in ML algorithms? 2. *Why* is it unfair that Black patients are sicker, on average, than White patients with the same risk score? a. How might a luck-egalitarian explain this unfairness? b. How might a critic of luck-egalitarianism explain this unfairness? 3. What ethical questions should we ask at each stage of algorithm development to address the problem of unfair bias?	Question 1: Algorithmic unfairness is a diverse phenomenon; different algorithms are unfair for different reasons (e.g., because they are trained on unrepresentative data, or because they are designed to predict an inappropriate target variable). It is important to give an overview of this diversity before homing in on a particular example. Question 2: It is easy to report an intuition that something is unfair; it is harder to articulate why it is unfair. A common reaction to the healthcare algorithm studied in class is that it is unfair because it disadvantages Black patients on the basis of factors outside their control. While this response has some merit, it should be problematized. Some inequalities—including, perhaps, health inequalities—should be compensated despite being the result of people's free choices. Question 3: Asking the right ethical questions at the right time is an ethical skill that can prevent bad decision-making.

Materials

Key Philosophical Concepts:	● Bias ● Fairness/unfairness ● Luck-egalitarianism	*Bias* is distinguished from unfair bias at the start of the module. Bias in the sense used here (roughly, "discrimination") is also distinguished from statistical bias, a notion with which students may be more familiar. *Fairness* is discussed primarily in the case study. Students are asked to articulate their sense that the algorithm discussed is unfair. *Luck-egalitarianism* is introduced as one account of what makes certain inequalities unfair. It is problematized using insights from Elizabeth Anderson's influential article, "What Is the Point of Equality?"
Assigned Readings:	● Barocas, Hardt, and Narayanan. "Introduction" to *Fairness and Machine Learning.* ● Ledford. "Millions of Black People Affected by Racial Bias in Health-Care Algorithms." *Nature NEWS*.	*Barocas, Hardt, and Narayanan* provide an excellent overview of the topic. They discuss why we use ML algorithms, provide examples of unfairly biased algorithms, explain how bias can arise, and introduce some measures of fairness. As ethically minded computer scientists, the authors provide technical depth that students may be seeking. *Ledford* provides a brief, accurate, and readable summary of the healthcare algorithm that is discussed in class and of the study published in *Science* that revealed the algorithm's racial bias.

Implementation

Class Agenda:

1. Introduction: Ubiquity of ML algorithms, why we use them, the problem of unfair bias, terminology.
2. How bias arises in ML algorithms: problem formulation, dataset construction, deployment, feedback loops.
3. Case study: How and why a major healthcare risk-prediction algorithm is racially biased.

	4. Class activity/discussion. 5. Luck-egalitarianism and key objections. 6. "Relabeling" as a solution to the problem of racial bias seen in this case. 7. Ethical questions to ask at each stage of algorithm development; Brief introduction to model cards.	
Sample Class Activity:	Students are asked to discuss the following questions in groups of 2-3 and share their responses with the class during a subsequent large-group discussion: 1. Why might someone think it is *unfair* for Black patients to receive lower risk scores than their White counterparts in this context? 2. Why might someone think it is *not* unfair?	Discussion questions are framed keeping in mind that this material is sensitive. Students are not asked to state their opinions about whether it is unfair but rather are asked to think about what arguments can be given on either side. The purpose of this activity is to ask students to go beyond intuitions about fairness and unfairness and to articulate reasons for thinking something is fair or unfair. In the subsequent discussion, they are equipped with concepts for structuring such explanations.
Module Assignment:	Before class, students are asked to answer the following reading response questions: 1. What 1-3 things did you learn or take away from today's reading? 2. What was unclear, or do you have more questions about today's reading? After class, students complete a post-class quiz. They are asked to answer three questions: 1. Why might a committed luck-egalitarian think that it is *unfair* for Black patients to receive lower risk scores than their White counterparts when these risk scores are used to recommend enrollment in a high-risk care program? 2. Do you think the luck-egalitarian gives a compelling explanation of why this is unfair? Why or why not? 3. Imagine that the healthcare risk-prediction algorithm discussed in class is used to determine insurance premiums, rather than inclusion in a high-risk care program, and that patients who are considered to be at high risk are charged higher insurance premiums. Articulate *two* ethical questions that you would ask during the process of deciding whether this practice is fair.	Students in this course are required to complete pre-class reading response questions and post-class quiz questions for each class session. These assignments serve dual purposes: to incentivize completion of readings and attentiveness during class, and to reinforce learning. The *pre-class reading response questions* are open-ended. Question 2 is meant to prime students to think critically and independently about the material. The *post-class quiz questions* are meant to reinforce concepts and tools introduced in class. Question 1 asks students to articulate the luck-egalitarian explanation of why the healthcare algorithm discussed in class is unfair. Question 2 asks students to critically reflect on this luck-egalitarian explanation. Question 3 asks students to practice generating ethical questions similar to the ones they

		saw in class, this time in a new context.
Lessons Learned:	Student responses to this module were overwhelmingly positive. Students found the concept of luck-egalitarianism particularly useful for thinking about the ethical issues discussed. They also appreciated the overview of how bias can arise at different stages of algorithm development and enjoyed discussing a single example in depth. Pedagogical lessons learned: ● Students appreciate having philosophical concepts and theories to structure their thinking about questions that seem at first to be nebulous (e.g., "Why is *X* unfair?"), even when those concepts and theories are simultaneously problematized. ● Real-world examples, especially from current events, go a long way in ensuring that module content is engaging and memorable. ● This module was taught to a class of approximately 250 students in a tiered, lecture-style classroom, with approximately 80 additional students participating online. The course sessions do not normally include discussion. In this setting, small-group-based (2-3 student) discussions followed by large-group discussions worked well; however, it is useful to dedicate a few minutes early in the session to having the students determine their small groups and introduce themselves to their discussion partners.	