CS 109A: Entry for Module Repository

1. **Course Name:** CS 109A: Introduction to Data Science
2. **Course Level:** Upper-level Undergraduate
3. **Course Description:** "This course is the first half of a one-year introduction to data science. We will focus on the analysis of data to perform predictions using statistical and machine learning methods. Topics include data scraping, data management, data visualization, regression and classification methods, and deep neural networks. You will get ample practice through weekly homework assignments. The class material integrates 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
   5. communication – summarizing results through visualization, stories, and interpretable summaries" (Course_Description)
4. **Module Topic:** Algorithmic Fairness and Recidivism Prediction
5. **Module Author:** Heather Spradley
6. **Tags:** prediction [cs], fairness [phil], bias [both], machine learning [cs], discrimination [phil]
7. **Module Overview:** In this module, we discuss algorithmic fairness, focusing on the special case of fairness in recidivism prediction. The central case study for the module is COMPAS, a recidivism prediction tool that is used widely in the criminal justice system. In 2016, ProPublica published a piece arguing that COMPAS is unfairly biased against black defendants on the grounds that the tool's false positive rate for black defendants is higher than its false positive rate for white defendants. Northpointe, the company that developed COMPAS, responded by arguing that the tool is "racially neutral" because it is calibrated between races: any two individuals that receive the same score are equally likely to reoffend, regardless of race. After reconstructing and evaluating both arguments (and drawing on John Rawls' views about procedural fairness in *A Theory of Justice*), we consider more general questions about fairness in recidivism prediction. How, in general, might preexisting racial bias affect the performance of recidivism prediction tools based on machine learning? What can data scientists working on recidivism prediction problems do to help ensure that the systems they develop are fair? And should the criminal justice system be using recidivism prediction algorithms to make decisions in the first place?
8. **Connection to Course Material:** In this course, students learn how to build predictive models and consider various problems that interfere with the accuracy of these models, such as feedback loops. In the module, we consider how to develop predictive models that are both accurate and fair. We also challenge the idea that ensuring fairness requires sacrificing accuracy, particularly in the case of recidivism prediction.
9. **Module Goals:**

- Introduce students to the topic of algorithmic fairness, with a focus on fairness in recidivism prediction.
- Consider various ways in which predictive algorithms might be unfair, as well as how to develop fairer predictive algorithms.
- Equip students with philosophical tools to help them think more clearly about algorithmic fairness.

10. Key Philosophical Questions:

1. What is fairness and what features of a predictive algorithm make it fair?
2. What kinds of features of an individual would a fair algorithm take into account?
3. In what ways can data collection be done fairly or unfairly and how does that impact the fairness of the predictive model that was trained on that data?

11. Key Philosophical Concepts:

- Fairness
- John Rawls' "veil of ignorance" thought experiment
- Moral relevance and irrelevance
- Bias
- Discrimination

12. Assigned Readings:

• Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias" (ProPublica).	This piece by ProPublica initiated the debate about whether COMPAS is biased against black defendants. In addition to introducing students to one of the central arguments in that debate, the reading provides useful background about COMPAS and how it is used in the criminal justice system.

13. Class Agenda:

1. Overview.
2. Case study: the COMPAS recidivism prediction tool.
3. ProPublica's argument that COMPAS is unfair.
4. Philosophical concepts: fairness, moral relevance, John Rawls' veil of ignorance thought experiment.
5. Technical concepts: false positive rates and calibration.
6. Argument that COMPAS is fair (based on Rob Long's article "Fairness in Machine Learning").
7. Data and data collection as further objections to the fairness of COMPAS.
8. Discussion.

14. Sample Class Activity: In order to get students to feel the force of the ethical questions about predictive algorithms used in recidivism prediction, the module begins with two polls. In the first poll, students are asked to consider a scenario in which they are a judge

making a pre-trial decision: they must decide whether to make that decision based on their own judgment or based on a risk assessment produced by a predictive algorithm. In the second poll, they are asked to make the same decision but from the perspective of the detainee about whom the pre-trial decision is being made. After the polls are complete, the students discuss their reasons for their answers. Then, they are asked to consider the common assumption that predictive tools will allow us to pass the buck on certain kinds of responsibilities in high stakes cases. They then discuss the way in which the responsibility of the creator of the predictive algorithm is heightened because of the way in which the algorithms are relied upon.

15. **Module Assignment:** In a post-module assignment, students are asked to explore recidivism data and corresponding COMPAS scores published by ProPublica. They are then asked to: (1) find correlations and differences between a defendant's race and various other variables in the data; (2) write a short response to the question, "With respect to these variables, how could bias in the data or data collection be impacting or causing these differences?"; (3) build three predictive models from the data that leave out race and other correlating variables in different ways in order to see what impact different variables are having on the model; and (4) discuss the resulting false positive rates amongst different racial groups in each of their models and what implications this has for the fairness of predictive algorithms.
16. **Lessons Learned:** Since the course had recently covered the technical concepts of calibration and false positive rates, we assumed that spending time reviewing these concepts would be unnecessary. In practice, however, we found that some students were not fluent enough with these concepts to readily apply them in the context of a new discussion about algorithmic fairness. When we teach the module again, we plan to spend more time reviewing these concepts before introducing new material.