Repository Entry
Embedded EthiCS @ Harvard Teaching Lab

Overview

Course: CS 1: Great Ideas in Computer Science

Course Level: Introductory Undergraduate

Course Description: "An introduction to the most important discoveries and intellectual paradigms in computer science, designed for students with little or no previous background. Explores problem-solving and data analysis using the Python programming language; presents an integrated view of computer systems, from switching circuits up through compilers and object-oriented design. Examines theoretical and practical limitations related to unsolvable and intractable computational problems, and the social and ethical dilemmas presented by such issues as software unreliability, algorithmic bias, and invasions of privacy."[1]

Module Topic: Algorithmic Fairness

Module Author: Krupa K. Appleton

Semesters Taught: Spring 2022

Tags: Algorithm [CS], bias [phil], algorithmic bias [both], implicit bias [phil], explicit bias [phil], accuracy [CS], fairness [phil], procedural fairness [phil], substantive fairness [phil], distributive justice [phil], egalitarianism [phil], proxy [CS], runaway feedback loop [CS]

Module Overview: In this module, we discuss how algorithms may be biased and in what sense, if any, algorithmic decision-making may be considered unfair. Students are introduced to the philosophical distinction between procedural and substantive fairness and taught to apply these concepts to different kinds of decisions made by algorithms. Much of the discussion, as well as the follow-up assignment, asks students to identify bias at work in real-world examples of algorithmic decision-making and then to articulate what makes the resulting decisions fair or unfair.

Connection to Course Material: One of the goals of this course is to introduce students to algorithmic thinking. This module brings students to consider ethical implications of using algorithms to make decisions that disparately impact different groups of people and in doing so may help to ameliorate or to exacerbate unjust societal distributions of goods and ills.

The nature of this course, with the expansive scope of topics surveyed, lends itself to a range of modules. Previous modules for the course focused on privacy (see, for example, https://embeddedethics.seas.harvard.edu/classes/cs-1-2021-spring). The Embedded EthiCS TA chose the topic of algorithmic fairness because (1) it comes up frequently in ethical debates inside and outside the computer science field, making it relevant to both CS and non-CS majors, (2) it connected to the work students in this course had completed leading up to the module date, and (3) it

[1] https://embeddedethics.seas.harvard.edu/classes/cs-1-2019-spring

		was an apt springboard for introducing a philosophical distinction (between procedural and substantive fairness) that students could apply to a broad range of CS and non-CS issues.

	Goals	
Module Goals:	1. Identifying two senses in which algorithms may be biased (namely by inaccurately representing the world or by accurately representing an unjust world). 2. Articulating two notions of what makes algorithmic bias unfair by reference to the philosophical distinction between procedural and substantive fairness. 3. Explaining the ways in which algorithmic decision-making may contribute to exacerbating or remedying unjust societal distributions of goods and positions.	
Key Philosophical Questions:	1. What is bias? 2. Is bias ever morally unproblematic? If so, under what conditions? 3. What is procedural fairness? Substantive fairness? 4. (How) can a decision be procedurally fair but substantively unfair? (How) can a decision be procedurally unfair but substantively fair? 5. In what sense, if any, is algorithmic bias unfair? 6. What would a more egalitarian distribution of societal goods and ills look like?	Much of the philosophical content of the module revolves around introducing philosophical concepts and distinctions and familiarizing students with how to apply them using hypotheticals and case studies. A future module for a more advanced course or with more time could build on this content by having students critically think about the philosophical material, such as by considering how to weigh procedural fairness against substantive fairness considerations.

	Materials	
Key Philosophical Concepts:	● Implicit and explicit bias ● Morally irrelevant characteristics ● Procedural fairness and substantive fairness ● Egalitarianism	● The TA introduced the distinction between implicit and explicit bias for the purpose of helping students to see that most instances of algorithmic bias are implicit bias. However, many students seemed to get stuck on this distinction at the expense of focusing on the other features of bias that were more directly relevant to the philosophical

		content. The TA would recommend briefly introducing this distinction verbally but not making it a core part of the module content. ● The TA only briefly introduced egalitarianism, for the purpose of giving content to the concept of substantive fairness. She chose this theory because she felt it would be an intuitive theory of justice to understand for students who were not familiar with the concept (and, indeed, they seemed to readily grasp it in class and on the homework assignment).
Assigned Readings:	● Pre-class reading (assigned in full): Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica, *Machine Bias* (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing	The pre-class reading for this module primed students on the notions of algorithmic bias and fairness and also motivated the lesson by introducing them to a recent, high-stakes, real-world case study of algorithmic unfairness. In this piece, ProPublica describes a recidivism risk assessment algorithm, called COMPAS, which has been deployed in myriad criminal justice decision-making processes across the United States. It then discusses research ProPublica undertook that revealed bias in the algorithm against black defendants. By engaging individuals who have been affected by COMPAS, the piece humanizes the issue of algorithmic bias by bringing students to see its downstream consequences on individuals' life prospects and outcomes. Students were given the following questions to think through as they completed the reading: 1. What are the benefits of a recidivism algorithm? That is, what are reasons a society may choose to use an algorithm to make criminal justice decisions?

		2. What are the risks and drawbacks of relying on an algorithm to make these decisions? 3. What bias and fairness concerns are raised by using such an algorithm? We will analyze the concepts of bias and fairness in depth during the lesson, but I want you to trigger your current intuitions about what they mean.

	Implementation	
Class Agenda:	1. Overview. 2. Priming activity (see below "Sample Class Activity"). 3. Stakes of algorithmic bias for societal distribution of goods and ills (jobs, housing, etc.). 4. Key concepts: bias (explicit versus implicit; morally problematic versus morally unproblematic; bias that results from inaccurately representing the world versus bias that results from accurately representing an unjust world) and fairness (procedural versus substantive) 5. Applying key philosophical concepts to the topic of algorithmic bias. 6. COMPAS case study, to apply concepts and lessons learned.	
Sample Class Activity:	At the beginning of the module, students are presented with the following case and divided into groups of 3-4 for discussion. The class then re-convenes for a full-class discussion to bring out intuitions about fairness and foreshadow concepts and distinctions that will be taught during the lesson. "Suppose that the age at which someone starts computer programming is strongly correlated with future success as a software engineer at Google. On average, boys tend to start computer programming earlier than girls. Would it be unfair for Google to use the age at which someone starts computer programming as part of their basis for deciding which software engineers to hire? Why or why not?"	The module used several small-group-based short active-learning exercises ("check-ins") to stimulate student engagement. We have found that such exercises help dramatically in keeping students engaged, and they worked particularly well in a class of this size (about 30 students).
Module Assignment:	Students received the following prompt:	The homework assignment had students apply the concepts and distinctions introduced in class to

	Please read this piece on Amazon's same-day delivery service: https://www.bloomberg.com/graphics/2016-amazon-same-day/. Then, please respond to the following two questions with answers of 1-2 paragraphs each. I do not expect you to do any outside research, though I encourage you to connect to lecture materials and the pre-class reading where relevant. Question 1. Some people think that Amazon's *process* for determining which neighborhoods would receive same-day delivery was unfair, but others disagree. What do you think? Explain your reasoning. Consider what role bias and proxies played in the process. Question 2. Imagine that you are a community organizer from Roxbury and that it is still excluded from same-day delivery service (it no longer is, following publication of the piece). Setting aside the issue of procedural unfairness, present your strongest case for why Amazon's approach was *substantively* unfair. Consider what societal goods and ills are at stake, given the nature of the service that is being unequally distributed. Bonus question: Identify a potential runaway feedback loop that might be generated by Amazon's approach to determining which areas get free same-day delivery service.	the case of a service with which they were likely familiar (Amazon) in an area close to campus (Roxbury, MA). This was meant to show them that the kinds of issues we talked about in class are live and relevant and to help them articulate positions on these issues with precision and nuance developed from exposure to philosophical tools. In formulating their responses, many students introduced considerations that we had not discussed in class (e.g., efficiency and costs). It may be helpful to make explicit what kinds of considerations should be weighed for purposes of the assignment to help focus the responses on content learned from the module itself.
Lessons Learned:	Student response to this module was overall very positive. A few lessons stand out: 1. Students were not as familiar with the COMPAS case as the TA anticipated they might be while developing the module. She opted for using that as the main case study because it is a case that they will benefit from being well-versed in for purposes of future discussions of algorithmic bias, in school and beyond. However, the TA would recommend requiring students to submit answers to guiding questions on the reading ahead of time. Having made the questions	

optional, the TA was not sure how many students had completed the reading and, accordingly, how much to introduce the material during the lesson before diving into discussion.

2. Responses to the homework assignment suggested that, on the whole, students developed a strong understanding of the distinction between procedural and substantive fairness. However, it was not always clear that they were able to distinguish between a decision being biased and a decision being unfair. A future module could elaborate further on the relationship between bias and fairness.
3. Students expressed being particularly interested in the distinction between procedural and substantive fairness and in understanding how an algorithm could be unfair despite being accurate. Some students also expressed wishing we could have broached the topic of how to actually address algorithmic unfairness going forward. The TA deliberately sidestepped this question during the module both to ensure that we could cover all the requisite concepts in sufficient depth and because the answer transcends the fields of both philosophy and CS, requiring enough time to generate and apply all the relevant considerations (tied to ethics, policy, industry constraints, etc.). A future module may consider making space for discussion of this question as a way to wrap up the module, or perhaps incorporate it into a short-response homework question.