CS 182 Repository Entry
Embedded EthiCS @ Harvard Teaching Lab

Overview

Course:	CS 182: Artificial Intelligence	
Course Level:	Undergraduate	
Course Description:	"Artificial Intelligence (AI) is already making a powerful impact on modern technology, and is expected to be even more transformative in the near future. The course introduces the ideas and techniques underlying this exciting field, with the goal of teaching students to identify effective representations and approaches for a wide variety of computational tasks. Topics covered in this course are broadly divided into problem solving, multi-agent systems, reasoning with uncertainty, and machine learning. Special attention is given to ethical considerations in AI and to applications that benefit society."	
Module Topic:	Thinking Responsibly About AI Systems	
Module Author:	Eliza Wells	
Semesters Taught:	Fall 2021-2022	
Tags:	responsibility [phil], stakeholders [phil], respect [phil], justice [phil], artificial intelligence [CS]	
Module Overview:	This module provides tools to help students cultivate personal responsibility when engaging with AI systems. It does so by helping students practice how to a) recognize stakeholders who will be affected by particular AI systems; b) understand different ways those stakeholders can be impacted by considering the lenses of benefits/harms, respect, and justice; and c) identify different points in AI systems design where interventions can improve impacts for stakeholders: data, design, and deployment. All of these concepts are illustrated by working through a real-life case study.	This is an interactive module that helps students cultivate skills by asking them to practice thinking through each step with each other and the Embedded EthiCS TA.
Connection to Course Material:	This course provided a broad, introductory level overview of AI systems. Students had learned a variety of technical tools for how to build AI systems. This module came at the end of the semester and so took a step back to consider AI systems generally rather than delving into one specific issue. The module encouraged students to bring the different tools they've learned in the class to bear on recognizing and addressing ethical problems.	This particular course had two previous ethics-related lectures on fairness and value alignment. Students had already encountered case studies about discriminatory loan systems, COMPAS, and self-driving cars, and discussed technical solutions to these problems. Since students were primed to think about ethics in particular AI problems, this module sought to step back and give them tools for thinking about ethics more generally.

Goals

Module Goals:	1. Cultivate positive responsibility by introducing tools for ethical decision-making 2. Understand the ethical lenses of benefits/harms, respect, and justice

	3. Consider different levels of intervention into AI systems 4. Apply these tools to case studies	
Key Philosophical Questions:	1. What are the moral responsibilities of computer scientists working on artificial intelligence? 2. Who is impacted by AI systems? 3. What are different ways in which they can be affected? 4. What choices can computer scientists make when building AI systems that impact stakeholders?	This module aims to help students answer the first question by reflection on the others. The module draws upon the ACM code of ethics to argue that computer scientists have a moral responsibility to make the world a better place for those who live in it, so identifying ways that AI systems harm stakeholders is to identify a space where computer scientists have a moral responsibility to mitigate those harms where they can.

Materials		
Key Philosophical Concepts:	● negative vs. positive responsibility ● stakeholders ● benefits/harms ● respect ● justice	The module uses the distinction between negative responsibility (who deserves blame when things go wrong?) and positive responsibility (how can I be aware of the impacts of my decisions?) to set up the importance of considering different stakeholders and ethical lenses in order to be positively responsible. Benefits/harms, respect, and justice as presented as distinct lenses that can assess different ethical dimensions of systems and situations. Students are reminded that these values can be in conflict: sometimes we have to make difficult decisions between, for example, systems that benefit more people and systems that are just.
Assigned Readings:	● Virginia Eubanks, "A Child Abuse Prediction Model Fails Poor Families," excerpt from *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* On Wired.com https://www.wired.com/story/excerpt-from-automating-inequality/	This article presents the module's central case study: the Allegheny Family Screening Tool, which is a predictive machine learning algorithm that seeks to assess risk of child abuse or neglect to determine whether investigation is needed. Eubanks discusses various failings with the AFST. The reading prepares students to think about different kinds of stakeholders and impacts that AI systems can have.

Implementation

Class Agenda:

1. Computer scientists and negative vs. positive responsibility
2. Case study: the Allegheny Family Screening Tool
3. Thinking responsibly about the AFST
    a. Who will be impacted by this system?
    b. How will they be impacted?
        i. Benefits/harms: What are the potential consequences of this system for each stakeholder?
        ii. Respect: How does this system show respect for each stakeholder's autonomy (think: transparency, consent, control, etc.)?
        iii. Justice: Does this process treat each stakeholder fairly? Does this process lead to fair outcomes?
    c. What technical choices influence these impacts?
        i. Data
        ii. Design
        iii. Deployment
4. Stepping back: are there other questions we should be asking in this process? What about an additional choice point: do or don't?

Sample Class Activity:

Students were asked to work through the case study with the Embedded EthiCS TA by discussing each step of the thought process in groups and then sharing what they discussed with the class.

The goal of the module was to help students practice the skill of recognizing ethical dimensions in real-life cases. The module was very interactive so students had to practice engaging with each other and thinking through the case study in real time.

Module Assignment:

There was no assignment for this module. A successful assignment for this class would give students more practice applying the thought process presented in the module.

One assignment option could have been presenting a different case study (perhaps a technology that has not yet been completed, such as self-driving cars or autonomous lethal weapons) and asking students to write a short essay or answer a series of questions that went through the process above for that case study.

Lessons Learned:

1. Students were engaged throughout and were able to bring different technical concepts from the course to bear on the case study.
2. The module would have benefitted from more examples of the kinds of impacts discussed at each step so students would

	have a model for their responses in the group.