Repository Entry Template
Embedded EthiCS @ Harvard Teaching Lab

Overview

Course:	CS 252r
Course Level:	Graduate
Course Description:	"Seminar course exploring recent research in programming languages. Topics vary from year to year. Students read and present research papers, undertake a research project. Fall 2020: We will explore programming languages for artificial intelligence. Programming Languages drive the way we communicate with computers, including how we make them intelligent and reasonable. In this advanced topic course, we will look at artificial intelligence broadly construed from the point of view of programming languages. We gain clarity of semantics, algorithms and purpose. Topics include differentiable programming, neuro-symbolic systems, constraint and probabilistic programming, interpretable AI and more. Reading and discussion will be based on a selection of papers, suggested collectively. Grading is based on participation, presentation and final project."[1]
Module Topic:	Interpretability and the Right to an Explanation
Module Author:	Zachary Gabor
Semesters Taught:	Fall 2020
Tags:	Interpretability [phil], Right to explanation [phil], Procedural fairness [phil]
Module Overview:	This module focuses on the relationship between the purported right to an explanation regarding automated decisions (decisions informed by or based entirely on algorithmic predictions) and fairness in machine learning. The topic is pitched in response to criticism of the EU's General Data Privacy Regulation (GDPR)'s protection of this purported right, which claims that focusing on interpretability is not the most promising means of achieving equitable results in such decisions. The module aims to familiarize students with considerations of fairness in decision-making which reach beyond issues of demographic discrimination. The in-class portion consisted of a presentation on Sunstein's two conceptions of procedural fairness in decision-making (rule-bound decisions vs. decisions based on individualized consideration) and an exercise in which students consider why fairness in particular decisions might demand one or the other conception of fairness.
Connection to Course Material:	An assigned reading for the course was Doshi-Velez and Kim's "Toward a Rigorous Science of Interpretable Machine Learning," in which they offer a taxonomy for evaluating interpretability. They propose various measures that can be used to assess The course is designed as a research-oriented topics course for graduate students. Students are invited to set the course agenda by selecting papers from a collection on the course website

[1] http://pl-ai-seminar.seas.harvard.edu/

	the quality of an explanation, depending on the sort of decision that is in question.	to present on. The assigned paper explained why the right to explanation is an important consideration for automated decision-making.

Goals		
Module Goals:	1. Introduce students to features of procedural fairness beyond non-discrimination. 2. Acquaint students with relationships between these features of fairness and features of explanations of decisions, including in the context of automated decisions. 3. Apply these tools in thinking about what counts as a satisfactory explanation in the contexts of different decisions.	
Key Philosophical Questions:	1. What features of a decision, besides discrimination on the basis of membership in a marginalized identity group, might contribute to the unfairness of a decision? 2. How do opportunities for human override of computer-generated decisions further fairness? How do they threaten fairness? 3. How can rigid conformity to rules in decision-making promote fairness? How can it be deleterious to fairness?	The goal of the module is to introduce students to the idea that explanations of important decisions have ethical importance over and above their usefulness in checking for errors in automated decisions. The key philosophical issues to broach are the ways in which a decision may be unfair even if it is not discriminatory on the basis of identity (Question 1), and how different ways of balancing between rigidity and opportunities for *ad hoc* adjustments can contribute to, or detract from, fairness (Questions 2 and 3).

Materials		
Key Philosophical Concepts:	● Discrimination ● Procedural fairness ● Fairness as individual consideration ● Fairness as rule-bound consistency	This module focuses on the notion of procedural fairness (under each of the two specified conceptions) as a matter of being given due consideration or "a fair shake."
Assigned Readings:	● https://www.openrightsgroup.org/blog/machine-learning-and-the-right-to-explanation-in-gdpr/ ● "Two Conceptions of Procedural Fairness" Sunstein, 2006.	The blog post summarizes discussion of the purported protection of a right to explanation in the EU GDPR, including an expression of skepticism about whether ensuring such a right is an effective way to limit discrimination. Sunstein's article

		introduces two notions of procedural fairness which can be used to illuminate ways in which a decision may be unfair without being discriminatory in the colloquial sense.

Implementation

Class Agenda:	1. Discussion: fairness beyond non-discrimination a. GF distinguishes between cases of unfairness which involve identity-based discrimination and other varieties of unfairness, for example the unfairness of decisions that are unjustly capricious. 2. Presentation: Sunstein's two conceptions of procedural fairness a. Discussion of Sunstein's distinction between fairness in decision making as rule-bound-ness and fairness as individualized consideration 3. Discussion: relationship between Sunstein's two concepts and measures of interpretability a. Discussion of the relationship between local and global interpretability, and of presence or absence of opportunities for human override in machine-aided decisions and the ways in which these affect fairness of both kinds. 4. Activity: analyzing automated decisions regarding bail	The presentation in class introduces by example the idea that a decision may be unfair without discriminating on the basis of identity simply by being capricious. It goes on to summarize the idea (from Sunstein) that sometimes fairness suggests that a decision be governed by uniform rules and, at other times, that a decision should be individualized to the particular case.
Sample Class Activity:	Students are split into two groups and asked to consider what they would want out of a decision-making regime for setting a criminal defendant's bail, one from the point of view of the government, and the other from the point of view of the defendant. Students are asked to discuss what they would want to know about how the decision was made in order to be comfortable accepting it, and about whether it mattered whether the decision was individually tailored and whether it mattered whether the decision was made according to consistently applied rules. Subsequently the whole class re-convenes to discuss.	The aim here is to get the students to distinguish between what the subject of a decision might want out of a decision and that to which they are entitled. This should emerge from the larger group discussion. The example is useful in that the two parties have competing interests, so even if technological barriers are not an issue, there are competing concerns to balance in thinking about how the decision is to be made. There are benefits and disadvantages to be weighed: from the point of view of the defendant vs that of the state; of making decisions more individualized vs more rule-bound; allowing vs

		disallowing human intervention; and so on. A key point to raise during this module is the question of *to whom* an explanation is owed for this kind of decision. Do due process requirements and related restraints on government authority require that a decision be justifiable to the defendant or to society at large? How does our answer to that question affect the kind of explanation we think is required in this case? Students adopting the perspective of a defendant were asked to consider what they would *like* out of an explanation, rather than what they were *entitled* to as an explanation. Likewise, students acting as the decision-makers were asked to discuss what they would like to know instead of what they think a defendant or another interested party is entitled to know. Doing so provides students a more concrete, less theoretical way of beginning to think through the relevant issues. In group discussion, we can proceed from the question of what kind of explanation might be desired to the question of what kind of explanation might be owed.
Module Assignment:	Students are asked to articulate a contrast between decision-contexts in which different requirements constrain what kind of explanation is called for, e.g. in handing down a judicial opinion as opposed to in directing traffic. The text of the assignment provides examples of the kind of principle that might be called for.	In the version I ran, students submitted these as forum posts, which enabled them to respond to one another's ideas.
Lessons Learned:	Students were engaged and interested in the topic, asking good, hard questions. They had surprising intuitions about the situations in which a subject is and isn't entitled to an explanation. Many seemed to reason that there was a positive correlation between 'high stakes' decisions and the amount of explanation required. In a future iteration, the Embedded EthiCS TA might raise examples in which it makes sense to sacrifice some degree of fairness in	

	order to avoid disastrous consequences, e.g. in organizing an evacuation in advance of a natural disaster. In general, frontloading more discussion of the differences in constraints across different situations may help.	