**Introduction to Embedded EthiCS Module for CS252r Fall 2020**

In our Embedded EthiCS module on the 29th, we will be discussing the relationship between fairness and the right to an explanation in ML-aided decision-making. We’ll be thinking about three different use cases: approving or denying loan applications, triaging patients in the emergency room, and granting or denying arrestees release on bail. In each case, we’ll be asking what a decision-maker might reasonably want out of explanations of ML-aided decisions, what the subject about whom the decision is being made might reasonably want, and what both parties could reasonably agree that the subject was entitled to.

**Readings**

In addition to the Doshi-Velez and Kim paper, we’re posting two readings, one required and one optional, on the website.

The required reading is a short blog post which presents an overview of some of the principal commentaries on the right to an explanation in tech ethics, in the context of the EU’s General Data Protection Regulation GDPR.

**Two Questions about the Reading to Mull Over Before Class**

Ruiz cites an argument to the effect that guaranteeing a right to an explanation of ML-aided decisions is unimportant, because doing so is not the most effective way to prevent discrimination. This argument presupposes that the only interest subjects of decisions have in a right to an explanation is that granting it is a guard against discrimination. But that is not obvious. For what other reasons might it be an a subject’s interest to have access to an explanation of a ML-aided decision about their life?

Ruiz cites a distinction between model-centric and subject-centric explanations, one which shares some affinity with Doshi-Velez and Kim’s distinction between global and local explanations. As the subject of a ML-aided decision, why might you be interested in getting a global explanation of the way the system makes decisions? Why might you be interested in getting a local explanation of how the decision about your particular case was made?

The optional reading is pp. 1-18 of Sunstein’s journal article “Two Conceptions of Procedural Fairness,” which discusses the importance to legal theory of two different ways of thinking about what it means for a decision-making regime to be fair. In class, I’ll start us off with an overview of the framework Sunstein provides and suggest some basic ways in which it is relevant for thinking about decision regimes that rely in whole or in part on ML-aided algorithms.

**Note** Sunstein’s article is a discussion of the ways in which notions of fairness can be applied to assess whether a decision-making regime passes legal (and more specifically Constitutional) muster. Legal theory and ethics are of course related in all sorts of ways, but they are not the same subject, so bringing Sunstein’s paper to bear will require a bit of transposition from the context “is this legal?” to the context “is this ethical?.” Part of the
work of my preliminary presentation on Tuesday will be directed toward this transposition, but I wanted to flag it as something to bear in mind if you do the optional reading.

Looking forward to our discussion!

Zach Gabor

G5 (Philosophy)

Graduate Fellow, Embedded EthiCS, AY 2020-21