CS 279R. Repository Entry
Embedded EthiCS @ Harvard Teaching Lab

Overview

Course: CS 279R Research Topics in Human-Computer Interaction (HCI)

Course Level: Graduate

Course Description: “This year only: Students will read and discuss HCI papers about computers working with---or clashing against---the strengths and weakness of human cognition, e.g., the positive and negative impacts of AI recommendation systems and the impact of interruptions on continuity of thought. Activities will include a small number of lectures, discussion of relevant literature in each field, and a project, in which students will work together in groups to design and carry out HCI research.”[1]

Module Topic: Ignorance & Algorithms

Module Author: Elís Miller Larsen

Semesters Taught: Fall 2020

Tags: Systems [CS] algorithmic design [CS] HCI [CS] unknown unknowns [CS] riskiest risks [CS] MNIST images [CS] predictive models [CS]/[phil] ignorance [phil] bias [phil] discrimination [phil]

Module Overview: This module examines and explores the ethical considerations for the new Turing test in AI research and HCI: unknown unknowns. Unknown unknowns are aspects of algorithmic design that we do not know, and do not know that we do not know. These information gaps are of ethical concern because they have the potential to generate problematic bias. The module breaks down the two ways that ignorance can impact algorithmic design. The first is when the system is ignorant, i.e., it doesn’t recognize that its own output is false. And the second is when a designer is ignorant, i.e., has a particular blind spot or exhibits some kind of bias. The module focuses on key cases when unknown unknowns generate ethical problems such as bias or discrimination in algorithm outputs. The aim is to better equip students to identify how unknown unknowns might arise for their own research projects.

Connection to Course Material: The module extends work that students have already done in the course regarding how to quantify uncertainty and identify riskiest risks in algorithmic design.

The topic is a timely addition, connecting to issues in HCI and AI research. It connects the problem of bias in algorithmic design to unknown blindspots. This illuminates one of the challenges faced by designers: anticipating and preventing problems without even being able to see the source of those problems . Previous outlines for identifying riskiest risks relied on the understanding that designers are able to

[1] https://canvas.harvard.edu/courses/74718/assignments/syllabus

		recognize and quantify ignorance. The existence of unknown unknowns puts pressure on this assumption.

	Goals	
Module Goals:	1. Recognize the problem of unknown unknowns for algorithmic design and identify why the problem generates ethical concerns. 2. Understand that the problem of unknown unknowns is not only at the individual level but also at the social level. 3. Differentiate between psychological and system ignorance. 4. Identify the pros and cons for proposals to mitigate ignorance.	The module identifies bias and discrimination as ethical concerns that system designers must be cognizant about. The goal of the module is to help students identify the part of the system design where bias can infiltrate a seemingly innocuous design. For example, bias can occur at the level of individuals in the form of implicit bias or other prejudices. Bias can also form via accidental, or unintentional connection between categories, such as race, gender etc. The main takeaway of the module is that students must be aware that bias, or ignorance, can go unnoticed, so designers must have practices and policies that can act as a safeguard against more harmful forms of ignorance, which can cause unjust outcomes.
Key Philosophical Questions:	1. How does ignorance generate bias within automated decision-making systems? 2. Is bias a form of ignorance? Is ignorance a form of bias? 3. Are we ethically responsible for designer or system ignorance?	The aim of these questions is to have students think about the role of ignorance in algorithm design and to think about whether such ignorance requires moral responsibility.

	Materials	
Key Philosophical Concepts:	● (Bayesian) Predictive Models ● Ignorance ● Bias ● Discrimination	HCI is a CS topic ripe for discussion on ignorance, bias, and discrimination. These concepts are critical for CS students to master as they impact how human subjects influence system design. Ignorance, bias and discrimination can be taught without direct reference to predictive models, but for this module, predictive models were utilized in order to show how predictive systems, as automated systems, can generate ignorance,

		bias, and discrimination in the outputs they generate, or by missing key hypotheses that might bear on the predictive feature.
Assigned Readings:	● Lisa Gitelman, *Raw Data is an Oxymoron*, Ch. 1 ● Safiya Umoja Noble, *Algorithms of Oppression*: *how search engines reinforce racism*, Ch. 1	*Raw Data is an Oxymoron* challenges the assumption that data is unbiased or “raw”. This reading prepares students for the idea that data might be infiltrated by biases and ignorance that will then impact system outcomes. *Algorithms of Oppression* provides specific examples of how ignorance, bias, and discrimination impact automated systems and the kind of harmful results that designers would want to avoid.

	Implementation	
Class Agenda:	1. Review ignorant systems and designer ignorance 2. Classify ignorance: known unknowns and unknown unknowns 3. Explain why ignorance leads to system error, bias, and/or discrimination 4. Activity: identifying unknown unknowns 5. Proposals for mitigating ignorance in automated systems	Marginal notes
Sample Class Activity:	*Phone a Friend*: Students already had a project group of 4-5 students where they were developing papers and proposals about problems in HCI. This activity is designed to work in tandem with this already assigned group project. Each group was paired with another group and had the task of helping each other recognize the unknown unknowns of their projects that they may not be able to see for themselves. Afterwards the class re-groups and we reviewed the kinds of unknown unknowns, or blind spots that groups came up with.	This activity is designed specifically for CS279 in order to work well with the final project. To modify this activity, instructors can still have students discuss the kinds of unknown unknowns that might impact an automated system. These would include things that a system might overlook, for example, in designing an autonomous vehicle, a designer might overlook having the system differentiate between a white t-shirt and a white plastic bag in the road. One would indicate a pedestrian and the other a benign piece of trash. Students can come up with examples and consequences in groups, and then come together to share with the entire class.
Module Assignment:	There was no specific assignment for this module.	Marginal notes [required]

Lessons Learned:	One thing that required clarification for this lecture is the relation between ignorance and system errors. It is important to be clear that when a system is ignorant there is also a system error, but not all system errors are instances of ignorance. In this module we discussed two ways that ignorance can impact system design. The first is that designers can be ignorant by exhibiting bias or prejudice that influenced the design. The second is that systems can be ignorant by making errors, or a system's inability to recognize an error. The third, which was not discussed in this module but could be added is consumer ignorance. Sometimes consumers do not know what they need or want from a system, and this can affect how well a system works for consumers.	Marginal notes