Ethical Algorithm Design

CIS 423/523

Prof. Michael Kearns

Spring 2022

Intros: MK

- @ Penn since 2002 (!)
- Spent 90s @ Bell Labs
- Interests:
  + machine learning/AI
  + algos/theory
  + trading/finance
  + game theory/econ
  + comp. social science
- Trading/Wall St. experience
- Tech consulting, Amazon

Intros: TAs

- Ira Globes-Harris
- Elizabeth Margolin

What's This Course About?

- What could go **wrong** with AI, ML, algos, etc.?

- **How** and **why** do things go **wrong**?

- **What** might we do to **fix them**?

At its core, a course on **algorithm design**, largely related to **AI/ML**

Course Background

- Previous pilots as 399
- Now "official"
- 423 vs. 523
- Satisfies SEAS engineering ethics requirement for:
  ASCS, BE, CMPE, CSCI, DMP, NETS

The Context

- Online/mobile revolution

- We've created massive
  digital trails of our
  interests, needs, locations,
  activities, health, habits,
  friends, family, hopes, fears...

- AI/ML/algos applied
  for personalization
  and prediction

- What could possibly
  go wrong?

The Problems

- Bias/discrimination
  + "unfair" models/algos
- Privacy leaks/breaches
  + e.g. your med. record
- Lack of explanability
  + why were you denied loan?
- Vulnerability to attack
  + e.g. self-driving cars

:

An Assertion

- For the most part, these problems are **not** the result of human malfeasance or incompetence

- Rather are **expected consequences** of the **standard principles** of AI/ML

So... we need to change these principles.

What would this even look like?

Algorithmic Fairness

- What does/should "fair" mean?

- Fair to whom?

- Group or individual?

- How to enforce?

Preview: Constrain ML training to obey fairness, study trade-offs

Algorithmic Privacy

- What does/should "privacy" mean?

- Breaches vs. leaks

- Breaches: crypto

- Leaks: ?

Preview: Anonymization is bogus; "right" notion of privacy involves randomization

Explainability

- Explanation of **what?**

- Training algo? Model? Specific decisions?
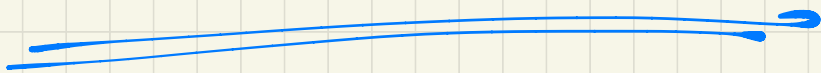
- How should an explanation "look"?

Preview: Very nascent, but ideas from game theory, linearization, elsewhere

Robust AI/ML

- Prevent attack/manipulation of algos/models
- E.g. change output by changing input
- E.g. data poisoning

Preview: Early days, but ideas for adversarial training, robust models

Mechanics & Materials

Resources & Comms

- Course website: main resource for videos, notes, readings, assignments, announcements

- Course Slack workspace for discussion, interaction & collaboration

- Occasional MK email

Prerequisites/Background

- **Required:** some basic programming (110 equiv.)

Useful:

- ML, data analysis
- stats/prob.
- algos/theory
- optimization
  ⋮

Readings

- Mainstream/general media articles

- Scientific articles

- General-audience books

- Web demos

Lectures

- Attendance required
- Participation encouraged & rewarded
- Discussions
- Guest lectures
- Demos

Assignments
(under construction!)

- Regular quizzes

- Coding assignments

- Midterm/final (?)

- Final projects or
  book club (?)

Next up:

Foundations of ML