CS 182: Ethics, Public Policy, and Technological Change

Rob Reich
Mehran Sahami
Jeremy Weinstein
Ece Korkmaz (Co-Head TA)
Alena Smith (Co-Head TA)
Today’s Agenda

1. The Technologist’s Mindset: Optimization (and its Discontents)
2. Introduction to Algorithmic Decision-Making
3. Algorithms and Fairness
4. What is Fairness?
   i. Fair process
   ii. Fair outcome
5. John Rawls and Abebe Birhane on Justice/Fairness
SURVEY RESULTS

What ethical questions and concerns have you experienced at work?

- Company funding war in Russia
- Making technology for private prisons
- Misrepresenting the capabilities of a system
- Military / weapons applications
- Working with authoritarian governments as clients
- Increasing user engagement / addiction for the sake of profits
- Working for a gambling client
- Selling ads for ethically questionable products
- Misleading language around people’s data collection
- Fear of speaking out about ethics due to job insecurity / retaliation from superiors / forced to listen to anti-union training
- Sharing data and videos about children
- Sexism, racism, and discrimination on the jobs
- **Positives:** Mandatory data privacy training, making products that made people’s lives easier, working on non-data centric applications
### THREE LEVELS OF ETHICS

<table>
<thead>
<tr>
<th>Level</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal Ethics</td>
<td>What does it mean to be a good person? What is your moral compass?</td>
</tr>
<tr>
<td>Professional Ethics</td>
<td>What are the standards of behavior that govern my professional role? What regulates responsible conduct as a member of a profession?</td>
</tr>
<tr>
<td>Social and Political Ethics</td>
<td>How to create rules and laws that facilitate well-being, cooperation, and human rights? Governance, law, regulation, and policy.</td>
</tr>
</tbody>
</table>
What’s Distinctive about Ethical Inquiry?

- Normative questions and conclusions; not about explanation
- Prescriptive, not descriptive
- Not about understanding the world as it is, but as it ought to be
Feynman’s (and Birhane’s) Error

• Feynman’s error was in viewing ethics as a one-time, already completed task. He engaged in ethical reasoning in 1941, reached a conclusion, and checked the ethics box as “complete.”

• What’s the remedy?

• Stop ethical sleepwalking. Awaken ethical ambition.

• Birhane and Feynman’s error: one reason why ethical reasoning should be an ongoing process is because our conception of what is ethical keeps changing.

• Big conclusion: Fairness is not a mathematical but a social concept.
Optimization Mindset
The Engineer’s Error?
Misunderstanding Optimization

It’s better to get something done efficiently rather than inefficiently.

It’s better still to optimize.

“Everything is an optimization problem”

Conclusion: efficiency/maximization/optimization are good, even intrinsically good.
“Everything is an optimization problem.”

Source: Wired Magazine, January 30, 2013; Flickr CC BY 2.0
Optimization becomes an orientation to life.
“In positioning itself as tech’s moral compass, academic computer science belies the fact that its own intellectual tools are the source of the technology industry’s dangerous power. A significant part of the problem is the kind of ideology it instills in students, researchers, and society at large. It’s not just that engineering education teaches students to think that all problems deserve technical solutions . . . [R]ather, the curriculum is built around an entire value system that knows only utility functions, symbolic manipulations, and objective maximization.”

-- Jimmy Wu, Optimize What?
What’s the Value of Efficiency?

Efficiency/maximization/optimization are **derivative** or **second-order** values.

They are instrumentally good.
Not intrinsically or fundamentally good.

Optimization is an instrument to some end or goal. We must independently assess the worthiness of that end or goal.
Speed bumps

Delayed results until Polls close

Deliberate in the jury box

The Architecture of Evil
On the delusion that technical work is morally neutral
Roger Forseger

Agricultural History
Volume IV  April, 1930  Number 2
ECONOMIC EFFICIENCY AND COMPETITIVE ADVANTAGES OF SLAVERY UNDER THE PLANTATION SYSTEM
By L. C. Gray

Sources The New Atlantis Summer 2012; Agricultural History by LC Gray, April 1930)
Three Problems of the Optimization Mindset

1. The Problem of Bad Ends/Goals/Objectives
2. The Problem of Finding Measurable Proxies for Good Goals
3. The Problem of Multiple Valuable Goals
Problem 1: means over ends

The Problem of Bad Ends/Goals/Objectives

• Doing something efficiently is only a means to accomplish an end (or an objective).

• From a moral standpoint, what’s important is the goodness or badness of end, not the means to accomplish it.

• Obsessing about efficiency can lead to optimizing for bad ends. And thereby making the world worse, not better.

• Also, many technologies are dual purpose: have good and bad use cases.
2. The Problem of Finding Measurable Proxies for Good Goals

What should an engineer optimize?

Need something *computationally tractable* and *representationally adequate* – a quantifiable measure of what we are trying to accomplish.

But goals like “happiness” or “safety” or “connecting people” are not easy to measure.
Facebook’s Mission Statement

Mission
Give people the power to share and make the world more open and connected.
So we connect more people

That can be bad if they make it negative. Maybe it costs a life by exposing someone to bullies. Maybe someone dies in a terrorist attack coordinated on our tools.

And still we connect people.

The ugly truth is that we believe in connecting people so deeply that anything that allows us to connect more people more often is *de facto* good. It is perhaps the only area where the metrics do tell the true story as far as we are concerned.
Problem 3: Multiple Valuable Goals

3. The Problem of Multiple Valuable Goals

Optimizing for one (good) end can threaten other valuable goals.
The Soylent Story

From Silicon Valley Powdered Meals to International Complete Nutrition Platform.
“I started wondering why something as simple and important as food was still so inefficient, given how streamlined and optimized other modern things are.”

Rob Rhinehart, Co-Founder
What’s Wrong With Soylent? (P.S. It’s not the taste)

About a week and a half ago, I began drinking Soylent every day. I can’t recommend that you do the same. For a purported breakthrough with such grand plans for reshaping the food industry, I found Soylent to be a punishingly boring, joyless product. From the plain white packaging to the purposefully bland, barely sweet flavor to the motel-carpet beige hue of the drink itself, everything about Soylent screams function, not fun. It may offer complete nourishment, but only at the expense of the aesthetic and emotional pleasures many of us crave in food.

Source: Farhad Manjoo, NY Times, May 28, 2014
The Multiple Purposes of Food

Bodily nutrition  Social connection  Gustatory Pleasure  Cultural identity

Optimize for one alone, and you have a problem.
1. What’s so good about efficiency and optimization?
2. Introduction to Algorithmic Decision-Making
3. Algorithms and Fairness
4. What is Fairness?
   i. Fair process
   ii. Fair outcome
5. John Rawls and Abebe Birhane on Justice/Fairness
The Great Promise of Algorithmic Decision-Making

If computer algorithms can more accurately and more efficiently outperform human at certain kinds of problem-solving, then algorithmic decision-making has enormous potential for good.


“This is the Big Data economy, and it promised spectacular gains. A computer program could speed through thousands of résumés or loan applications in a second or two and sort them into neat lists, with the most promising candidates on top. This not only saved time but was also marketed as fair and objective. After all, it didn’t involve prejudiced humans digging through reams of paper, just machines processing cold numbers.”
Algorithms and Values

So what’s the connection between algorithms and values like fairness?

If an algorithm is a mathematical function to translate inputs into a determinative output, what do values like fairness have to do with math or logic?

What does fairness have to do with $2 + 2 = 4$?

Let’s take an example of some immediate relevance to you...
CS182 Grading Rubric

Assignments and Grading Breakdown

In addition, the course includes five assignments. You will receive more information about each of the assignments well in advance of their due dates.

- Technical assignment – **due January 26 at 11:59pm PST**
- Philosophy paper – **due February 9 at 11:59pm PST** [NOTE: WIM students will have an additional revision due **February 25 at noon PST**]
- Group policy assignment – **due March 2 at 11:59pm PST**
- Technical assignment – **due March 9 at 11:59pm PST**
- Final reflection paper – **due on March 16 at 11:59pm PST**

Grades will be calculated as follows:

<table>
<thead>
<tr>
<th>Non-WIM Students</th>
<th>WIM Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation – 20%</td>
<td>Participation – 15%</td>
</tr>
<tr>
<td>Technical Assignment 1 – 14%</td>
<td>Technical Assignment 1 – 14%</td>
</tr>
<tr>
<td>Philosophy Paper – 20%</td>
<td>Philosophy Paper (original) – 9%</td>
</tr>
<tr>
<td>Policy Assignment – 20%</td>
<td>Philosophy Paper (revision) – 21%</td>
</tr>
<tr>
<td>Technical Assignment 2 – 6%</td>
<td>Policy Assignment – 20%</td>
</tr>
<tr>
<td>Final Reflection – 20%</td>
<td>Technical Assignment 2 – 6%</td>
</tr>
<tr>
<td></td>
<td>Final Reflection – 15%</td>
</tr>
</tbody>
</table>
Possible CS 182 Grade Schemes

• Grading on curve: pre-determined number of As. Grades are determined on the basis of performance in class relative to other students in class.

• Grading according to an external and objective standard. Everyone can earn an A, in principle.

• Grading relative to your own past performance: that you show effort and improve over the quarter factors into your grade.
Still More CS 182 Grade Schemes

- Everyone gets an A, assuming you show up and submit work. No relationship to actual quality of work.

- Grades are a function of hours worked on class assignments (monitored by Stanford IT, Zoom, and Canvas). More work = higher grade. Person who works the most gets the highest grade.

- People who visit my office hours get higher grades.
What about an algorithm?

Thought Experiment

Your school is concerned about biases in grading. So they train an algorithm to predict students’ final grades from demographics, prior performance, etc. It’s called GPAbot.

So even through instructors assess students’ assignments, GPAbot is ultimately what suggests a final letter grade for each student to instructors.

Do you endorse this? Why?

Thought experiment courtesy of Nick Byrd via twitter.com.
What Values are at Stake in the CS 182 Grading Algorithm?

1. **Fairness**: does the grading scheme treat people fairly?
   
   NOTE: multiple reasonable conceptions of fairness

2. **Transparency**: is the grading scheme made transparent? (or is it a black box?)

3. **Privacy**: are grades made public? [FERPA violation!]

4. **Due Process**: is there a fair process for appeal?

5. **Other Values**:
   - Justice: are grades the right way to assess students?
   - Power: are CAs legitimate graders?
Promise and Peril of Algorithmic Decisionmaking

Algorithms used widely, across some of the most important domains of life.

**Promise**: improve accuracy and efficiency over human decisionmaking. Humans make mistakes, and they’re slow.

Diminish or eliminate bias, randomness, human fallibility.

**Peril**: widespread adoption encodes bias, threatens fairness, privacy, transparency, and due process.
CASE STUDY

ALGORITHMIC DECISION-MAKING AND ACCOUNTABILITY
For Discussion

- Imagine that Stanford (or any other selective university) proposes to admit undergraduates on the basis of an algorithmic model
  - Train the model on past applications and student admissions.
  - What would be the benefits and risks of using automated review of applications?
Stanford Apologizes for Limiting Jewish Admissions in the 1950s

The university also issued a report that found that some administrators had “regularly misled” people who raised concerns about anti-Jewish bias.
Today’s Agenda

1. What’s so good about efficiency and optimization?
2. Introduction to Algorithmic Decision-Making
3. Algorithms and Fairness
4. **What is Fairness?**
   i. Fair process
   ii. Fair outcome
5. John Rawls and Abebe Birhane on Justice/Fairness
What is Fairness?

Substantive vs. Procedural Fairness

SUBSTANTIVE FAIRNESS = measured by the outcomes

PROCEDURAL FAIRNESS = focus on fairness of the process

Note: Fairness applied to Individuals and to Groups
Outcome Fairness/Substantive Fairness

Fairness involves equal treatment and expectations of fair outcomes.

Three examples:

• Orange juice experiment

• Ultimatum game

• Monkey experiment
Orange Juice Experiment, Inspired by Cornell Economist Robert Frank
Evolutionary Fairness: Ultimatum Game
An Evolutionary Sense of Fairness?

Frans de Waal, Moral Behavior in Animals, TEDxPeachtree Talk
Is Fairness always Equal Treatment?

No: Substantive fairness might require unequal treatment!

Let’s think about school funding...
Substantive Fairness

Lesson: **Substantive Fairness is CONTEXTUAL and DEPENDENT on a social understanding of morally relevant and irrelevant characteristics of persons.**

Aristotle:
Fairness is “treating likes alike and unlikes unalike”

Core Q: In what ways are people relevantly alike, and in what ways relevantly unalike?

This is not a question of math....
Procedural Fairness

Q: Can a fair process deliver a fair outcome?
PROVABLY FAIR SOLUTIONS.

Spliddit offers quick, free solutions to everyday fair division problems, using methods that provide indisputable fairness guarantees and build on decades of research in economics, mathematics, and computer science.
Spliddit's tasks calculator fairly divides household chores, work shifts at a hospital, or any other set of tasks. You begin by providing a list of tasks that you wish to assign (for example, morning shift, afternoon shift, night shift) and a list of participants. We then send the participants links where they specify how much they prefer each task relative to the others. Our algorithm uses these evaluations to propose a fair division of the tasks among the participants.

**Fairness Properties**

**Equitability**

The assignment of tasks is equitable if all participants believe their workload is identical.
A Fair Procedure?

Perfect procedural justice
A Fair Procedure?

2. Pure procedural justice
What about Historical Injustice?

"You do not take a person who, for years, has been hobbled by chains and liberate him, bring him up to the starting line of a race and then say, 'You are free to compete with all the others,' and still justly believe that you have been completely fair.

Thus it is not enough just to open the gates of opportunity. All our citizens must have the ability to walk through those gates."
Algorithmic fairness is not the only goal

Birhane, *Algorithmic Injustice, A Relational Ethics Approach*

“...we seek to center the needs and welfare of those that are disproportionately impacted and not solutions that benefit the majority. Most of the time this means not simply creating a fairness metric for an existing system but rather questioning what the system is doing, particularly examining its consequences on minoritized and vulnerable groups.”

**Structural vs. Incremental Reform**
Today’s Agenda

1. What’s so good about efficiency and optimization?
2. Introduction to Algorithmic Decision-Making
3. Algorithms and Fairness
4. What is Fairness?
   i. Fair process
   ii. Fair outcome
5. John Rawls and Abebe Birhane on Justice/Fairness
Rawls’s Justice as Fairness: Basic Framework

• Domestic justice, not global justice
• Society is a cooperative venture for mutual advantage
• A theory of distributive justice
• Justice as a pure procedure
• Provide an alternative to utilitarianism (which “fails to take seriously the distinction between persons”)

[Reference to domestic justice, cooperative venture, distributive justice, and alternative to utilitarianism]
How to Select Principles of Justice?

• The correct principles are those that all persons would rationally agree to.
• “principles that free and rational persons concerned to further their own interests would accept in an initial position of equality as defining the fundamental terms of their association” (10).
Problem: the choice situation?

• How shall we generate conditions that permit rational self-interest but do not reflect underlying inequalities, and differential bargaining power?
• Different natural talents among people
• Different inherited positions among people
• ➔ What could neutralize these inequalities in the choice situation?
Rawls’s Answer: The Original Position

“The idea of the original position is to set up a fair procedure so that any principles agreed to will be just.”

-Rawls, p. 118

This is why he calls his theory “Justice as Fairness”
Rawls’s Answer: The Original Position

- In the Original Position, you do not know:
  - your place in society
  - your natural assets or abilities
  - your class position or social status
  - your intelligence or strength
  - your race or your sex and gender identity
  - your conception of the good or what ends you will pursue in life.

- You are situated behind a “veil of ignorance”
A Fair Choice Situation

“This ensures that no one is advantaged or disadvantaged in the choice of principles by the outcome of natural chance or the contingency of social circumstance. Since all are similarly situated and no one is able to design principles to favor his particular condition, the principles of justice are the outcome of a fair agreement” §3, p. 11.

=>> People are free and equal under these conditions
Concluding Questions: Algorithmic Fairness

• How should we approach the idea of algorithmic fairness, and avoiding algorithmic bias? Substance or process? Individual or group?

• What to do if there is a trade-off between predictive accuracy and fairness, transparency, privacy, explainability, due process?

• How does algorithmic decisionmaking reinforce or contest existing power dynamics in the world?