CS 182: Ethics, Public Policy, and Technological Change

Rob Reich
Mehran Sahami
Jeremy Weinstein
Hilary Cohen
Today’s Agenda

1. Navigating trade-offs
2. Mapping policy and legal approaches
3. Making them work: technical considerations
4. Making them work: social scientific considerations
5. Should we be satisfied with tweaking the system?
New York City Council

OPEN THE CODE

epic.org/algorithmic-transparency
The proposed Algorithmic Accountability bill would:

- Require city agencies to publically release the source code of all algorithms they use to make their decisions
- Allow members of the public to “self-test” the algorithms by submitting their own data and getting the results

In committee, the bill was watered down to:

- Establish a task force that would examine exactly what algorithms are in use and whether any appear to discriminate
- It would also make recommendations on how to help New Yorkers understand algorithms and challenge their results.
The NYC task force recommended that NYC:

- Establish org. structure and guidance for overseeing use of ADS; provide resources and support to agencies
- Broaden public discussion of ADS
- Formalize agency reporting on ADS, establish single point of inquiry, create process for assessing harm

The critics argue that more ambition was needed including:

- Process for public consultation before ADS is designed or acquired
- Clear standard and process for regularly assessing disprop. impact, responding if discovered (incl. right of private action)
- Formal procedures for explanation of decisions within 20 days
- Wide-ranging transparency requirements
Put Yourself in Their Shoes

Now imagine you are the Mayor having received the task force recommendations. What position would you take?

• Would you accept the recommendations as is? Why or why not?
• If not, what would you prioritize in terms of strengthening the recommendations?
• Why would you prioritize these things?
• What gives you confidence that these mechanisms would be effective?
Criteria

1. Does it work? (Efficiency to achieve public safety)
2. Is it fair? (Fairness)
3. Can people understand how it works? (Transparency)
4. Can people appeal its judgment? (Due Process)
5. Does it use information that an individual might reasonably expect to remain private? (Privacy)

Today we are going to focus on how we balance the value of an algorithm in terms of achieving public safety and other values.

It means asking: What else do we value? How do encode that in law or policy?
Today’s Agenda

1. Navigating trade-offs
2. Mapping policy and legal approaches
3. Making them work: technical considerations
4. Making them work: social scientific considerations
5. Should we be satisfied with tweaking the system?
Substantive fairness is context-dependent. It depends on some *social understanding* of what fairness requires.

- That’s a hard problem to solve. But it isn’t a technical problem. It demands engagement in politics.

Procedural fairness is distinct from substantive fairness. We can envision what we might want a process to look like, especially from an “original position,” without knowing our own characteristics.

- This is what a lot of policy efforts are focused on. But the devil is still in the details.
Who Should be Governing?

If algorithms are doing (or have the potential to do) harm, who should we rely on to fix the problem?

- The engineers who create the algorithms
- The companies that hire the engineers
- The industries in which companies are organized
- The consumers who use the technologies
- The citizens who are impacted by new technologies
- The government which sets and enforces rules in the place where the new technologies are built/deployed

How does your answer differ depending on whether the algorithm is being used by a company or a government agency?
How confident are you in relying on companies to...

- To regulate themselves?
- To self-regulate as an industry?
- To be responsive to user pressure?
- To be responsive to their own employees?

**How dominant?**

Global market share
April 2018, %

- **Search**
  - Google: 91%
  - Smartphone web traffic:
    - Apple: 45%

- **Social media**
  - Facebook: 66%
  - Online retail:
    - Amazon: 37%

Source: Global Stats Counter

Economist.com
“Governments should intervene, at a minimum, when private action has negative public consequences; when shortsighted actions threaten to cause long-term harm; when failure to intervene undermines significant constitutional values and important individual rights; when a form of life emerges that may threaten values we believe to be fundamental; and when we can see that failing to intervene on the side of right will simply strengthen the interventions on the side of wrong.”

-- Larry Lessig, *Code 2.0*
Relying on Democratic Politics

The Myth of the Rational Voter
Why Democracies Choose Bad Policies

BRYAN CAPLAN
Maybe the Courts Will Save Us?

• The proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined
• Because COMPAS risk assessment scores are based on group data, they are able to identify groups of high-risk offenders – not a particular high-risk individual
• Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism
• A COMPAS risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed
• COMPAS was not developed for use at sentencing, but... for determinations regarding treatment, supervision, and parole
FAccT to the Rescue!

ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.
Three Critical Levers

Transparency.
What data is used? What results are generated? What is the underlying source code? What is the logic for the result?

Auditing.
Can independent experts test the algorithm? Who chooses them? What happens to their input? Is it made public?

Due process.
What results were generated for me? What data of mine was used? What if the algorithm is wrong? To whom can I appeal the algorithmic decision?
Running Into Tradeoffs

- Accuracy vs. fairness
- Accuracy vs. interpretability
- Accuracy vs. privacy
- Safety vs. transparency
- Safety vs. autonomy
- Etc...
Today’s Agenda

1. Navigating trade-offs
2. Mapping policy and legal approaches
3. Making them work: technical considerations
4. Making them work: social scientific considerations
5. Should we be satisfied with tweaking the system?
Algorithm Interpretability

- What makes an algorithm *interpretable*?
  - Ability to scrutinize source code?
  - Understanding results from algorithm?
  - Understanding general “logic” of the algorithm?

- How do we trade-off interpretability vs. accuracy?
  - Generally, more complicated model can be more accurate
  - Some models are more interpretable than others (e.g., a linear function, versus a set of rules, versus a neural network)
  - Hard/impossible for human to understand sufficiently complex models
    - It’s essentially a “black box” even if we have access to the code
    - Sometimes, we learn a simpler (or more readily understood) model to approximate the function in a complex model

- And now, it’s *Possibly Apocryphal Story Time*!
  - Show me the tanks!

- Neural network is weighting features of x-ray machine.
- Portable machines are more likely to be used with sick patients (who can't come in to the hospital).
- This creates a confound for trying to deploy real systems and can lead to misleading results.
Interpretability via Outcomes

- Accounting for outcomes (outcome-based explanation)
  - How particular inputs lead to particular output

- Legal requirement for credit scoring by Fair Credit Reporting Act (FCRA)
  - E.g., You have a right to know factors that impact your FICO score

- Adverse action notice
  - When credit denied you must be given specific reason for denial
  - Idea is to alert and educate the consumer about credit

- Can’t make credit decision based on protected attributes
  - But, proxy attributes can still be possible inputs (even in explanation)
Interpretability via Logic

• Logic of decision-making (logical explanation of result)
  • Need to provide description of “rules” in the system

• European General Data Protection Regulation (GDPR) requires access to “meaningful information about the logic involved” in automated decision-making that impact data subject
  • In case you were wondering, you are the “data subject”
  • More about GDPR to come

• Seems more transparent, but potentially more problematic
  • What is suitable “logic” for decision?
  • Can someone explain who can see my Facebook post if I post a picture tagging a friend who only allows her friends to see posts on her feed, but I let anyone see mine?
  • How about why Google showed me a particular set of search results?
Auditing Algorithms

- Basic idea of audit: field study to diagnose harmful discrimination/impacts from a decision-making process
  - E.g., Send the same resume with “male” or “female” name to see if there is a difference in interview rates

- Note: algorithms are embedded in software systems/platforms
  - In many cases, we are really auditing the platform, not just a single algorithm in it, which can create more complexity

- Auditing of algorithms/platforms can take different forms
1. Code Audit

- Platform makes underlying algorithm available to auditor
  - Could open source code (i.e., everyone is auditor)
  - Could put algorithm in safe escrow (i.e., designated auditors)

- Provides algorithmic transparency at source code level
  - Essentially, what question 1 of your assignment was about

- Doesn’t necessarily provide data algorithm was trained on
  - That’s why your assignment didn’t stop after question 1
  - Providing training data can be a privacy concern
    - We’ll talk much more about data privacy in the next unit
Some Machine Learning Architectures

- Accessing algorithm ≠ understanding how decisions are made
  - High-dimensional, non-linear, non-monotonic, discontinuous functions

\[ d(X) = w_1 X_1 + w_2 X_2 + w_3 X_3 + w_4 X_4 \]

Perceptron

```
   X_1  w_1
     |    |
     |    |
    X_2  w_2
     |    |
     |    |
    X_3  w_3
     |    |
     |    |
    X_4  w_4

Perceptron
```

“Simple” neural network

```
  X_1
   |
   |
  X_2
   |
   |
  X_3
   |
   |
  X_4

```

LeNet-5 Deep Neural Network architecture

```
INPUT 32x32
  C1: feature maps 6@28x28
  S2: f. maps 6@14x14
  C3: f. maps 16@10x10
  C5: layer 120
  F6: layer 84
  OUTPUT 10
```

LeNet-5 Deep Neural Network architecture
Proprietary Algorithms

• Algorithm can be core intellectual property of company
  • E.g., Email spam filter, Google ranking algorithm

• Release of algorithm could be untenable
  • Gaming/spamming of platform
  • Release of trade secrets to competition
  • Outright theft/copying
2. Noninvasive User Audit

- Auditors get users to give them results of using platform
- Generally, not randomized in condition assignment
  - Difficult to focus on what you believe may be discriminatory
- Sampling bias
  - Which users you have access to
  - Which users agree to share results with you
  - (Very) limited sample size
- Results from platform may be too sensitive in some domains
  - Finance, healthcare, criminal justice, etc.
3. Scraping Audit

- Auditors scraps results from platform using program/script
- Likely to violate US Computer Fraud and Abuse Act (CFAA)
  - Criminalizes unauthorized access to computer systems
- Likely to violate platform’s “Terms of Service”
- Should platforms provide an API (or modify their terms of service) to enable this?
  - How do you protect the platform from abuse?
  - Was Cambridge Analytica just auditing Facebook?
4. Sock Puppet Audit

- Auditors create “false users” (sock puppets) to interact with platform
  - E.g., Has been used to test if there is bias in Facebook ad targeting
- May violate US Computer Fraud and Abuse Act (CFAA)
  - Injecting “false” data into system (e.g., fake profiles in Facebook)
- Likely to violate platform’s “Terms of Service”
- Is this reasonable?
- Will it scale?
5. Crowdsourced Audit

- Auditors crowdsource real users to interact with platform
- Likely does **not** violate US Computer Fraud and Abuse Act
- Likely does **not** violate platform’s “Terms of Service”
- Does paying confederates change things?
- What if confederates were paid to click on ads?
Crowdsourced Auditing in Action

• TechCrunch, Jan 15, 2019
• Activists in Germany crowdsourcing credit scores (Schufa)
• OpenSchufa: project where users donate their financial data from Schufa
  • 3,000+ data donations

Canada has launched an Algorithmic Impact Assessment tool

- Scorecard to highlight issues in design and deployment of automated decision-making (ADM)
- "Unspoken expectation... that any government entity is to use this tool... to supervise, control, and mitigate potential issues with the deployment of an ADM system."

Available on Github: https://canada-ca.github.io/aia-eia-js/

Source: https://towardsdatascience.com/understanding-canadas-algorithmic-impact-assessment-tool-cd0d3c8cafab
Proposal From Microsoft President

- Brad Smith, President and Chief Legal Officer at Microsoft, has proposed a marketplace solution
- He claims, no one wants unfair algorithms
- Have companies evaluate their algorithms on a standard benchmark using a public dataset and report the results
- Consumers can choose to purchase/use the algorithm they deem best, given fairness results, price, etc.
- Do you think this is a reasonable proposal?
“Amazon scraps secret AI recruiting tool that showed bias against women”
-- Business News, Oct. 9, 2018

By 2015, the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way.

That is because Amazon’s computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

It penalized resumes that included the word “women’s,” as in “women’s chess club captain.” And it downgraded graduates of two all-women’s colleges...

The Seattle company ultimately disbanded the team by the start of last year because executives lost hope for the project...
1. Navigating trade-offs
2. Mapping policy and legal approaches
3. Making them work: technical considerations
4. Making them work: social scientific considerations
5. Should we be satisfied with tweaking the system?
The GDPR – A New Model

Better protection for personal data

- Clear consent required to process data
- Limits on the use of automated processing of data to make decisions, for example in the case of ‘profiling’
- Right to rectify and remove data, including the ‘right to be forgotten’ for data collected as a child
- Right to notification if data is compromised
- More and clearer information about processing
- Easier access to personal data
- Right to move data from one service provider to another
- Stricter safeguards for transfers of personal data outside the EU
The “Right to Explanation”

- Key GDPR provisions around algorithmic decision-making
  - A right not to be subject to algorithmic decision-making, except under specified circumstances
  - A right to obtain human intervention, to respond, and to contest any decision
  - A right to information about the data that is used, including one’s own personal information
  - A right to explanation about the logic involved in the decision
The Problem(s) with Transparency

It is simply one input into a complex decision-making process for individuals & groups.

Someone needs to pick up the information and run with it.

Who will do that, when, and why?
A ton of decisions are involved in making an auditing system work.

1. Who is the auditor?
2. How many are there?
3. Who selects them?
4. Who pays their salary?
5. Who do they report to and how?
6. Are there sufficient resources to get excellent auditors?
7. What happens to the findings of the auditor?
8. Who receives them?
9. What triggers a response and from whom?
“If men were angels, no government would be necessary. If angels were to govern men, neither external nor internal controls on government would be necessary. In framing a government which is to be administered by men over men, the great difficulty lies in this: you must first enable the government to control the governed; and in the next place oblige it to control itself.”

-- James Madison, Federalist Papers No. 51
Due Process for Whom?

Perceptions of how blacks are treated in the U.S. vary widely by race

% saying blacks are treated less fairly than whites in the country

- In dealing with the police: Whites 50%, Blacks 84%
- In the courts: Whites 43%, Blacks 75%
- When applying for a loan or mortgage: Whites 25%, Blacks 66%
- In the workplace: Whites 22%, Blacks 64%
- In stores or restaurants: Whites 21%, Blacks 49%
- When voting in elections: Whites 20%, Blacks 43%

Note: Whites and blacks include only non-Hispanics.
“On Views of Race and Inequality, Blacks and Whites are Worlds Apart”

PEW RESEARCH CENTER
Today’s Agenda

1. Navigating trade-offs
2. Mapping policy and legal approaches
3. Making them work: technical considerations
4. Making them work: social scientific considerations
5. Should we be satisfied with tweaking the system?
Minutiae vs. What Matters

Should we be satisfied with tweaking an algorithm to make it more fair or ensure transparency? Why not tackle the rigged system instead?