27: AI and Ethics

Jerry Cain
June 3, 2024

Lecture Discussion on Ed

What can be done
How AI is impacting our lives?

- Smartphones
- Social Media Platforms
- E-Commerce
- Autonomous Vehicles
- Security & Surveillance
- Navigation
- Banking & Finance Sector
- Smart Home
We live in a time with real work to be done. How can we begin to use ML to help with access to high quality education, smart grids, criminal justice reform, and better healthcare?
Facebook slammed by UN for its role in Myanmar genocide

https://www.cjr.org/the_media_today/facebook-un-myanmar-genocide.php


Bitcoin Devours More Electricity Than Many Countries

Annual electricity consumption in comparison (in TWh)

<table>
<thead>
<tr>
<th>Country</th>
<th>Electricity Consumption (TWh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>6,453</td>
</tr>
<tr>
<td>USA</td>
<td>3,990</td>
</tr>
<tr>
<td>Germany</td>
<td>524</td>
</tr>
<tr>
<td>All the world's data centers</td>
<td>205</td>
</tr>
<tr>
<td>Bitcoin</td>
<td>143</td>
</tr>
<tr>
<td>Norway</td>
<td>124</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>71</td>
</tr>
<tr>
<td>Switzerland</td>
<td>56</td>
</tr>
<tr>
<td>Google</td>
<td>12</td>
</tr>
<tr>
<td>Facebook</td>
<td>5</td>
</tr>
</tbody>
</table>

* Bitcoin figure as of May 05, 2021. Country values are from 2019.
Sources: Cambridge Centre for Alternative Finance, Visual Capitalist
Learning Goals

1. Understand limits in fairness through unawareness

2. Know two ways to measure fairness

3. Know some techniques to mitigate fairness issues
"... To call attention to the privacy risks, he [Michal Kosinski of Stanford’s GSB] decided to show that it was possible to use facial recognition analysis to detect something intimate, something ‘people should have full rights to keep private.’"

**Why Stanford Researchers Tried to Create a ‘Gaydar’ Machine**


"Presented with photos of gay men and straight men, a computer program was able to determine which of the two was gay with 81 percent accuracy, according to Dr. Kosinski and co-author Yilun Wang’s paper."

"'the algorithmic equivalent of a 13-year-old bully'"

"Indeed, few of the claims made by researchers or companies hyping its potential have been replicated, said Clare Garvie of Georgetown University’s Center on Privacy and Technology.

'At the very best, it’s a highly inaccurate science,' she said of promises to predict criminal behavior, intelligence and other character traits from faces. 'At its very worst, this is racism by algorithm.'"
How to Use ChatGPT and Still Be a Good Person

It’s a turning point for artificial intelligence, and we need to take advantage of these tools without causing harm to ourselves or others.


10 Ways GPT-4 Is Impressive but Still Flawed

OpenAI has upgraded the technology that powers its online chatbot in notable ways. It’s more accurate, but it still makes things up.


"We’re at the beginning of a broader societal transformation," said Brian Christian, a computer scientist and the author of... a book about the ethical concerns surrounding A.I. systems. "There’s going to be a bigger question here for businesses, but in the immediate term, for the education system, what is the future of homework?"

"OpenAI, the company behind ChatGPT, declined to comment for this column."

GPT-4 does drug discovery.

Give it a currently available drug and it can:

- Find compounds with similar properties
- Modify them to make sure they’re not patented
- Purchase them from a supplier (even including sending an email with a purchase order)

""We’re at the beginning of a broader societal transformation," said Brian Christian, a computer scientist and the author of... a book about the ethical concerns surrounding A.I. systems. "There’s going to be a bigger question here for businesses, but in the immediate term, for the education system, what is the future of homework?"

"OpenAI, the company behind ChatGPT, declined to comment for this column."

GPT-4 does drug discovery.

Give it a currently available drug and it can:

- Find compounds with similar properties
- Modify them to make sure they’re not patented
- Purchase them from a supplier (even including sending an email with a purchase order)
Philosophy and Ethics Ask Very Good Questions of CS

Here are a few questions and concepts worth discussing:
• What is a protected demographic?
• What is distributive harm? What is quality-of-service harm?
• What is fairness? How can various definitions of fairness be made core to machine learning?
Logistics Regression Is That Linear Separator

- Logistic regression computes some line that separates instances where $y = 1$ from those where $y = 0$.

\[ \theta^T x = 0 \]
\[ \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_m x_m = 0 \]

- We call such data (or the functions generating the data) linearly separable.
- Naïve Bayes is linear as well, since the different features are assumed to be conditionally independent.
## Frameworks of Harm

<table>
<thead>
<tr>
<th>Quality-of-service harm</th>
<th>Distributive harm</th>
<th>Existential harms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurs when a system does not work as well for one person as it does for another</td>
<td>Occurs when AI systems withhold opportunities, resources, or information</td>
<td>Occurs when AI gravely alters the course of all humankind</td>
</tr>
</tbody>
</table>

**Examples:**
- generative art
- facial recognition
- document search
- product recommendation

**Examples:**
- hiring
- lending
- college admissions
- salary and benefits

**Examples:**
- democracy
- climate
- genocide
- AI supremacy
Under-sampling, Lack of Data, Poorly Curated Datasets

Initial explanations for AI-driven harm:

- On both gender and race, majority groups are generally overrepresented in image databases.
- Most images in some widely used databases: white faces.
- **Faces In The Wild** database was *83.5% white and 77.5% male*.
- Machine learning may ignore or deemphasize features of minority groups.
Intentional Improvements in Face Datasets in 2018

Research and activism by Joy Buolamwini, Timnit Gebru, and many others has led to more representative datasets already.

"The Gender Shades project pilots an intersectional approach to inclusive product testing for AI"
- Gender Shades

Stanford PhD 2017

Stanford University
Algorithmic Discrimination: The Case of St. George’s Hospital

2,500 applicants to the medical school

Interview approx. 625 (so \(\frac{3}{4}\) are rejected)

Offer spots to approx. 425 (so 70% of interviewees accepted)

In 1979, Vice Dean Dr. Geoffrey Franglen completes a classification algorithm to do the job.
A computing professional has an additional obligation to report any signs of system risks that might result in harm. If leaders do not act to curtail or mitigate such risks, it may be necessary to "blow the whistle" to reduce potential harm. However, capricious or misguided reporting of risks can itself be harmful. Before reporting risks, a computing professional should carefully assess relevant aspects of the situation.
Algorithmic Discrimination: The Case of St. George’s Hospital

This biased result was predictable.

At least 60 people unfairly rejected each year.

1. Codifying misogyny and racism
   Previous admissions process was biased against female applicants and applicants of color. Simply learning from the data will replicate and perpetuate the past bias.

2. Improper use of sensitive features.
   Algorithm relied on data like name and place of birth that provide no information about the merit of the applicant and are highly correlated with sensitive categories like race and gender.

3. Can be biased without intention to be evil
   Even if you didn’t mean to make a biased algorithm, that doesn’t mean it isn’t biased.
Two Philosophical Views of Fairness

**Procedural Fairness:**
Focuses on the decision-making or classification process, ensures that the algorithm does not rely on unfair features.

**Distributive Fairness:**
Focuses on the decision-making or classification outcome, ensures that the distribution of good and bad outcomes is equitable.
Three Formal Definitions of Fairness

Fairness through Unawareness
Fairness through Awareness: Independence
Fairness through Awareness: Separation
Fairness through Unawareness

Motivating idea: "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race" – Chief Justice Roberts

Note: Fairness through unawareness of some federally protected categories—that is, a subset of sensitive features—is legally required in domains like lending.

How to do it:
1. Exclude sensitive feature (race, gender, age, etc.) from your dataset
2. Also exclude proxies to the sensitive feature (name, zip code)
Protected Demographics

Protected Groups
Protected groups under **EEO** are race, color, national origin, religion, age (40 or older), sex (including pregnancy, sexual orientation, or gender identity), physical or mental disability, and reprisal.

*Equal Employment Opportunity, US*

Similarly defined for housing, loans, etc.
Case Study: Facebook Ads & Job/Housing Recommendations

Facebook creates "Lookalike" feature for advertisers: upload a "source list" and find users with "common qualities" to target ads for goods and services, including housing and jobs.

March 2018: National Fair Housing Alliance (NFHA) & other civil rights groups sue Facebook over violations of the Fair Housing Act.

March 2019: As part of settlement, Facebook agrees not to use "age, gender, relationship status, religious views, school, political views, interested in, or zip code" in creating lookalike audience.

https://techscience.org/a/2021101901
https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias
Facebook Input Lookalikes

Create a Lookalike Audience

1. Select Your Lookalike Source
   - Select an existing audience or data source
   - Create New Source

2. Select Audience Location
   - Countries > North America
   - United States
   - Search for regions or countries

3. Select Audience Size
   - Number of lookalike audiences
     - 2.3M
   - Audience size ranges from 1% to 10% of the combined population of your selected locations. A 1% lookalike consists of the people in your lookalike source, increasing the percentage creates a bigger, broader audience.

Create a Special Ad Audience

1. Select Your Source
   - Select an existing audience or data source

2. Select Audience Location
   - Countries > North America
   - United States
   - Search for regions or countries

3. Select Audience Size
   - Number of Special Ad Audiences
     - 1
   - Audience size ranges from 1% to 10% of the combined population of your selected locations. A 1% Special Ad Audience consists of the most similar online behavior to your source. Increasing the percentage creates a bigger, broader audience.
New Special Ad Audiences Still Biased

gender: equally biased
age: almost as biased
race: more difficult to measure given the tools provided but still biased
politics: less biased

Figure 2: Gender breakdown of ad delivery to Lookalike and Special Ad audiences created from the same source audience with varying fraction of male users, using the same ad creative. We can observe that both Lookalike and Special Ad audiences reflect the gender distribution of the source audience, despite the lack of gender being provided as an input to Special Ad Audiences.

Two Philosophic Values of Fairness

**Procedural Fairness:**
Focuses on the decision-making or classification process, ensures that the algorithm does not rely on unfair features.

**Distributive Fairness:**
Focuses on the decision-making or classification outcome, ensures that the distribution of good and bad outcomes is equitable.

Fairness through unawareness
(Facebook example shows this isn’t always effective)
## Fairness Through Awareness Terms

- **$D$:** protected demographic
- **$G$:** guess of your model (aka $y$ hat)
- **$T$:** the true value (aka $y$)

### $D = 0$

<table>
<thead>
<tr>
<th></th>
<th>$G = 0$</th>
<th>$G = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 0$</td>
<td>0.21</td>
<td>0.32</td>
</tr>
<tr>
<td>$T = 1$</td>
<td>0.07</td>
<td>0.28</td>
</tr>
</tbody>
</table>

### $D = 1$

<table>
<thead>
<tr>
<th></th>
<th>$G = 0$</th>
<th>$G = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 0$</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$T = 1$</td>
<td>0.02</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Distributive Fairness #1: Parity

**Fairness definition #1: Parity**
An algorithm satisfies “parity” if the probability that the algorithm makes a positive prediction ($G = 1$) is the same regardless of begin conditioned on demographic variable.

$D$: protected demographic

$G$: guess of your model (aka $y$ hat)

$T$: the true value (aka $y$)

$$P(G = 1 | D = 1) = P(G = 1 | D = 0)$$
Distributive Fairness #2: Calibration

**Fairness definition #2: Calibration**
An algorithm satisfies “calibration” if the probability that the algorithm is correct ($G = T$) is the same regardless of demographics.

\[ P(G = T \mid D = 1) = P(G = T \mid D = 0) \]

- $D$: protected demographic
- $G$: guess of your model (aka y hat)
- $T$: the true value (aka y)
Distributive Fairness #2: Calibration (Relaxed)

**Fairness definition #2: Calibration**
An algorithm satisfies “calibration” if the probability that the algorithm is correct ($G = T$) is the same regardless of demographics.

\[
\frac{P(G = T | D = 1)}{P(G = T | D = 0)} \geq 1 - \epsilon
\]

where $\epsilon = 0.2$

US legal standard: "disparate impact" also known as the 80% rule.
What does fairness through awareness fail to capture?

- If the classifier is worse at identifying candidates (e.g., for an experimental surgery) in a minority group, the candidates selected might experience worse outcomes, leading to future bias.
- Quality-of-service disparity might lead to allocation disparity.
- Dwork et. al. (including Omer Reingold of Stanford) call this a "self-fulfilling prophecy".

[Link](https://dl.acm.org/doi/10.1145/2090236.2090255)
Advanced Idea: Adversarial Learning [aka: train bias out]

Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction

Christina Wadsworth
Stanford University
Stanford, CA
cwads@cs.stanford.edu

Francesca Vera
Stanford University
Stanford, CA
fvera@cs.stanford.edu

Chris Piech
Stanford University
Stanford, CA
piech@cs.stanford.edu

"Recidivism prediction scores are used across the USA to determine sentencing and supervision for hundreds of thousands of inmates. One such generator of recidivism prediction scores is Northpointe's Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) score, used in states like California and Florida, which past research has shown to be biased against black inmates according to certain measures of fairness. To counteract this racial bias, we present an adversarially-trained neural network that predicts recidivism and is trained to remove racial bias."
COMPAS: Predicting Recidivism

$X$

data about an inmate: their zip code, past crimes, etc.

$$\text{argmax } P(y|x)$$

$y = \{0, 1\}$

$\hat{y} = 0$

prediction whether they will commit a crime again
Can We Train Out Bias?

Model 1: Prediction  
\[ x \xrightarrow{\theta_1} \hat{y} \]

Model 2: Extract Demographic  
\[ \theta_2 \xrightarrow{} \text{Demographic} \]

\[ \theta_1, \theta_2 = \arg\max_{\theta_1, \theta_2} L_1(\theta_1) - L_2(\theta_2) \]

*note in the paper these were neural nets*
Can We Train Out Bias?

Before: COMPAS is Biased

After: Gaps are reduced

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

Accuracy

Parity Gap

Calibration Gap

Accuracy

Parity Gap

Calibration Gap
160,000,000,000,000 hashes per second

Climate change and bitcoin aren’t a significant part of ethics within Stanford CS yet.
It isn’t too hard to see the trend

We will most almost certainly hit 2x CO2 before 2060, and then blow past it.
Easy to Know Impacts Will Be Harsh

Without prompt, aggressive limits on CO₂ emissions, the Earth will likely warm by an average of 4°-5°C by the century’s end.

**HOW BIG A CHANGE IS THAT?**

In the coldest part of the last ice age, Earth’s average temperature was 4.5°C below the 20th century norm. Let’s call a 4.5°C difference one “ice age unit.”

-2 IAU

Snowball Earth (-4 IAU)

-1 IAU

20,000 years ago

Average during modern times

0 IAU

Where we are today

+1 IAU

Where we’ll be in 86 years

+2 IAU

Cretaceous Hothouse

+200m sea level rise

No glaciers

Palm trees at the poles

Lisa Yan, Chris Piech, Mehran Sahami, Katie Creel, and Jerry Cain, CS109, Spring 2024
Impacts are Here

Cyclone Idai
Impacted over 3M people
It is hard to feel like you can do anything...

"I am just going to wait and see what happens"
What Can We Do?: Push for some change

- Individual
- Community
- Nation State

Is this our sweet spot?
What Can We Do? Reduce CS "Pump" of Proof of Work

Ethereum’s Response?
"Ethereum switched on its proof-of-stake mechanism in 2022 because it is more secure, less energy-intensive, and better for implementing new scaling solutions compared to the previous proof-of-work architecture."

---

Lisa Yan, Chris Piech, Mehran Sahami, Katie Creel, and Jerry Cain, CS109, Spring 2024

Stanford University
What Can We Do? Advocate for a Clean Grid in CA

Source: California Energy Commission, staff analysis November 2018
Build?

Tech-For-Good Startups Recently Started By Stanford Graduates

- Recidiviz (Clementine Jacoby)
- Edlyft (Arnelle Ansong)
- Develop For Good (Mary Zhu)