Ethics Lecture
by Ecy
Housekeeping

- Assignment 4, Crypto is due tonight at 11:59 pm
  - Grace Period until Wednesday, August 2nd at 11:59 pm
- Midterm Scores are out on Gradescope!
  - Regrade requests due by Thursday at 1:30 pm
Today...

- **Ethics, Values, Examples & Consequences**
  - Why is ethics important?
  - Case studies and consequences
  - Defining bias, potential harms, fairness

- **Designing For Our Values**
  - Looking at Assignment 5
  - Problem framing and examining language
  - Ethics and image manipulation
  - Combatting bias, asking questions

- **Next Steps**
  - Further Steps & Resources
  - Ethics goals in CS106A
Why does ethics matter?
<table>
<thead>
<tr>
<th><strong>Human Dignity</strong></th>
<th><strong>Moral Guidance</strong></th>
<th><strong>Social Harmony</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ethics recognizes the inherent worth and dignity of every individual, promoting respect and consideration for their rights and well-being.</em></td>
<td><em>Ethics provides a framework of principles and values that guide individuals and societies in making moral decisions and resolving ethical dilemmas.</em></td>
<td><em>Ethical behavior fosters cooperation, trust, and empathy, leading to healthier and more harmonious relationships within communities.</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Responsible Decision-Making</strong></th>
<th><strong>Global Impact</strong></th>
<th><strong>Ethical Leadership</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ethical considerations help individuals and organizations make responsible decisions that take into account the consequences of their actions on others and the environment.</em></td>
<td><em>In a connected world, ethical decisions have far-reaching consequences. Acting ethically contributes to a more just and sustainable global community.</em></td>
<td><em>Ethical leaders inspire trust and motivate others to act responsibly, creating positive and ethical organizational cultures.</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Personal Growth</strong></th>
<th><strong>Justice and Fairness</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Embracing ethics can lead to personal growth and the development of strong character traits, such as integrity, empathy, and humility.</em></td>
<td><em>Ethics advocates for fairness and justice, aiming to treat all individuals equitably and impartially, regardless of their backgrounds or circumstances.</em></td>
</tr>
</tbody>
</table>
But how do these values relate to the world of code?
Imagine you want to create an algorithm that analyzes whether a baby will make a good US president.

And you feed it this (training) data
What do you think it will produce when asked to predict? Why?
Society can be biased.
Bias Syllogism

...and these biases can be reflected in data.

For example, take Machine Learning (ML), a subset of AI. It is designed to find patterns in (training) data and hook onto those patterns to make matching predictions.

Thus, ML can reinforce and even exacerbate societal biases.
This has consequences.
Goal

Train a model to automate Amazon recruitment
"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."
Amazon scraps secret AI recruiting tool that showed bias against women

Result:

"In effect, Amazon’s system taught itself that male candidates were preferable. It penalized resumes that included the word “women’s,” as in “women’s chess club captain.”

Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory, the people said."
"Some 55 percent of U.S. human resources managers said artificial intelligence, or AI, would be a regular part of their work within the next five years, according to a 2017 survey by talent software firm CareerBuilder."
Facial Recognition

Impact:
A lack of representation in the data can lead to technology not working as planned for certain groups.
In trying to predict recidivism rates "the formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants."
What is bias?
**Statistical Bias**
The difference between the measured results, or output, and the "true" value or expected result.

It is the mathematical meaning of bias.

**Discriminatory Bias**
Discrimination resulting from a negative attitude toward the social group (e.g. animus or indifference).

**Indirect Discrimination**
Discrimination that does not result from such an attitude, but from rules and procedures constructed in a way that favors one group over another.
Discriminatory Bias in Data

Discrimination as defined by the Stanford Encyclopedia for Philosophy:

"The rules and norms of society consistently produce disproportionately disadvantageous outcomes for the members of a certain group [and] the outcomes are unjust to the members of the disadvantaged group"

Biased measurement or classification

+  

Use of that bias that compounds existing injustice

=  

Discriminatory or Unfair Bias
Two Examples

a. Ratings for Uber drivers were found to be lower for "BIPOC" drivers. Drivers with too low of ratings would be fired.

b. Scores on a nursing licensing exam in the United Kingdom were statistically greater for women compared to men. Upon further review, it was found that women tended to perform better on questions about caring for a baby/infant.

Biased measurement or classification

+ Use of that bias that compounds existing injustice

= Discriminatory or Unfair Bias
What kinds of harm might this result in?
Representational Harm
A person is harmed when her identity is diminished in public representations of her social groups.

Who is represented in this data?
Who can see themselves in it?

Allocative harm
A person is harmed when opportunities resources, benefits, and protections that would otherwise be allocated to them are unfairly withheld.

What are greater implications of less allocation to a group?

Amazon Rekognition Performance on Gender Classification
98.7% 68.6% 100% 92.9%
DARKER MALES DARKER FEMALES LIGHTER MALES LIGHTER FEMALES
How might this affect fairness?
<table>
<thead>
<tr>
<th>Principle</th>
<th>Parity Premise</th>
</tr>
</thead>
<tbody>
<tr>
<td>The distribution of goods should be based on morally relevant characteristics, not on morally arbitrary ones.</td>
<td>Because we are equal, we should adjust rules and procedures to ensure that outcomes reflect that.</td>
</tr>
</tbody>
</table>

**Example:** People are equally likely to be a good teacher => expect numbers of highly rated teachers proportionate to population

*Note: VERY common metric of statistical fairness*
Fairness Definitions

Formal Equality of Opportunity

- Positions that confer great advantages should be open to all applicants.
- Applications are assessed on their relevant merits.
- Applicant deemed most qualified according to appropriate criteria is offered the position.

Example

Everyone has the same opportunity to develop skills needed for the job, apply for the job, and get promoted.

Substantive Equality of Opportunity

- Takes into account systemic inequalities to ensure everyone in a community has access to the same opportunities and outcomes.
- Acknowledging that inequalities exist and working to eliminate them.

Example:

Affirmative action: “Race-conscious, holistic selection processes are essential to achieve diversity in STEM programs at selective colleges and universities and to create a pipeline of diverse talent in STEM”

- Stanford amicus brief in 2022
Designing For Our Values
Assignment 5: Bias Bars

Looking at datasets
Showcasing Values through Design

How can we showcase our values through design and how we collect, use, and understand data?

Examine Problem Framing
How are we framing the problem we are going to solve?

Watching Language
How can language reinforce existing biases in data?
Programming is problem solving
How do we frame problems?

How things are framed can affect which solutions we pursue and fundamentally change the nature of how we solve what we may deem the "same" problem.

Problem: We need to get rid of people living on the streets...

Problem: Some people don't have a home!
Language and Data
Bias can pop up in the language people use to describe things.
### Descriptive vs Normative Language

There are different kinds of language we can use to describe things.

<table>
<thead>
<tr>
<th>Descriptive Language</th>
<th>Normative Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statements of fact</td>
<td>Evaluative language</td>
</tr>
<tr>
<td>What people did</td>
<td>Express opinions/reactions</td>
</tr>
<tr>
<td>What happened</td>
<td>How things should be</td>
</tr>
<tr>
<td><strong>How things are</strong></td>
<td><strong>right</strong>/<strong>wrong</strong></td>
</tr>
<tr>
<td>&quot;lectures are 75 minutes long&quot;</td>
<td>&quot;good&quot;/&quot;bad&quot;</td>
</tr>
<tr>
<td>&quot;sections are mandatory&quot;</td>
<td>&quot;should&quot;/&quot;should not&quot;</td>
</tr>
</tbody>
</table>
Thick Normative Language

Thick Normative Language = Descriptive + Normative Language Combined

Thick Normative

- express morally or aesthetically “loaded” descriptions

- Cowardly
- Cautious
- Polite
- Rude
- Chill

- Kind
- Caring
- Smart
- Knowledgeable
- Professional
Images and Manipulation
How do we get information?

How do we learn about things in the world?

- **Perception**
  - Direct from senses
  - ![Eyes]

- **Testimony**
  - Info from others
  - ![Speech bubbles]

- **Mathematical Deduction/Reasoning**
  - ![Magnifying glass]
## Harms from Image Manipulation

When we manipulate images, what is understood to be highlighting the truth, and what is understood to be a misrepresentation of it?

<table>
<thead>
<tr>
<th>Manipulation</th>
<th>What makes altered image trustworthy?</th>
<th>Damaging Image/Speaking for others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manipulation is hidden influence that subverts another person's decisionmaking power (Nissenbaum).</td>
<td>Modeler/illustrator should explain which idealizations have been made and for what purpose.</td>
<td>Image and audio manipulation can be used to make others appear to say or do things they did not say or do.</td>
</tr>
</tbody>
</table>
Next Steps...
DALLE-2 Searches

show me an oil painting of a computer science teacher

Computer Science Teacher Race

<table>
<thead>
<tr>
<th>Computer Science Teacher Race</th>
<th>Percentages</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>64.1%</td>
</tr>
<tr>
<td>Asian</td>
<td>13.4%</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>10.2%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>7.1%</td>
</tr>
<tr>
<td>Unknown</td>
<td>4.9%</td>
</tr>
<tr>
<td>American Indian and Alaska Native</td>
<td>0.3%</td>
</tr>
</tbody>
</table>
Efforts are being made!

Reducing bias and improving safety in DALL-E 2

Today, we are implementing a new technique so that DALL-E generates images of people that more accurately reflect the diversity of the world’s population.

In April, we started previewing the DALL-E 2 research to a limited number of people, which has allowed us to better understand the system’s capabilities and limitations and improve our safety systems.

During this preview phase, early users have flagged sensitive and biased images which have helped inform and evaluate this new mitigation.

We are continuing to research how AI systems, like DALL-E, might reflect biases in its training data and different ways we can address them.

During the research preview we have taken other steps to improve our safety systems, including:

- Minimizing the risk of DALL-E being misused to create deceptive content by rejecting image uploads containing realistic faces and attempts to create the likeness of public figures, including celebrities and prominent political figures.
- Making our content filters more accurate so that they are more effective at blocking prompts and image uploads that violate our content policy while still allowing creative expression.
- Refining automated and human monitoring systems to guard against misuse.

These improvements have helped us gain confidence in the ability to invite more users to experience DALL-E.
Trade-offs/ Values

Other Twitter users who tested DALL-E 2 replied to Woolf’s thread sharing the same issue – specifically regarding race and gender biases. They suspected OpenAI’s diversity solution was as simple as the AI’s appending gender- or race-identifying words to the user-written prompts without their knowledge to inorganically produce diverse sets of images.

“The way this rumored implementation works is it adds either male or female or Black, Asian or Caucasian to the prompt randomly,” Woolf said in a phone interview.
Combatting bias

Check for Statistical Bias
What correlations and patterns exist in my dataset? In what ways do they fail to accurately represent the world?

Check for Discriminatory Bias
In what ways do the biases compound existing injustice?

Decide how to use the data given bias
For what social purposes would it be appropriate to use this data? How should we communicate information about possible biases?
Questions to ask about fairness

Values in data set
- What conception of fairness is encoded in the data set, if any?
- Does it lead to discrimination?

Values in data-based decisions
- Given existing biases in the data set, would it be fair to rely on them for our decisions?
- Would decisions based on the data set lead to discrimination?
Examining our own
Who does our data include/exclude?
Self-Examination

The Article About the Play

‘106’ surprises audiences with a philosophical discussion of technology

[Image of three actors sitting around a table with coffee cups.]

Three actors sit around a table with the spotlight on them, performing in one of the vignettes from TAPS show, "CSI 06A." The cast included Aiyana Washington ‘24 (above, left), Sophia Wang ‘26 (center) and Peter Li ‘25 (right). (Photo: BRAD YAC-DIAZ/The Stanford Daily)
GI Joe Fallacy

"The G. I. Joe fallacy refers to the misguided notion that knowing about a bias is enough to overcome it (Santos & Gendler, 2014).

The name of this fallacy derives from the 1980s television series G. I. Joe, which ended each cartoon episode with a public service announcement and closing tagline, “Now you know.”"

Harvard Business School

An Example within Stanford's HAI
Tradeoffs

"Everyone cares about ethics in tech until they get their contract."

-CS182 lecture
Remember:

We are always Learning, 
Adapting, and Growing!
Some Additional Resources

Personal Class Recs

- **CS182**: Ethics, Public Policy, and Technological Change
- **PUBLPOL 103F**: Ethics of Truth in a Post-Truth World
- **CS 278**: Social Computing
- **SYMSYS 201**: Digital Technology, Society, and Democracy

Centers

- [McCoy Family Center for Ethics in Society](#)
- [HAI institute](#)
Ethics goals for CS106A

- Image manipulation should not compromise people’s autonomy
- End to end encryption addresses some privacy considerations
- When using data, especially big data, our choices our values
- Think about how our programming design decisions can affect others
## Recap

Today, we talked about...

### Ethics Context and Definitions
- Why ethics matters
- Examples of how ethics can show up in society
- Definitions of bias
- Looking at societal harm
- Looking at fairness
- Looking at ethics in the context of CS106A

### Implementation
- How does framing problems impact us?
- How can language contain bias?
- How can we help to combat biases in data?
- What're next steps we can take?
THE END