CS440/ECE448 Lecture 7: Fairness

Mark Hasegawa-Johnson

CC-SA 2.0, Kathy Simon, 2008

https://commons.wikimedia.org/wiki/File:Viola_and_Mina_share_food.jpg

Outline

- Fairness Problems
  - Opacity; Scale; Damage
  - Redlining
- Definitions of Fairness
  - Individual Fairness; Counterfactual Fairness
  - Demograpic Parity vs. Equal Odds vs. Predictive Parity
  - The 4^{th} box?

Benefits of Statistical Models

- Before statistical models, many decisions were blatantly unfair
  - College admissions: Who were your parents?
  - Housing loans: Does the loan officer like the way you look?
- In many cases, statistical models are provably more accurate and more fair
  - College admissions: Weighted sum of grades, SAT, essay, interview
  - Housing loan: Weighted sum of income, debt, education

Problems with Statistical Models

(Cathy O'Neil, Weapons of Math Destruction, 2016)

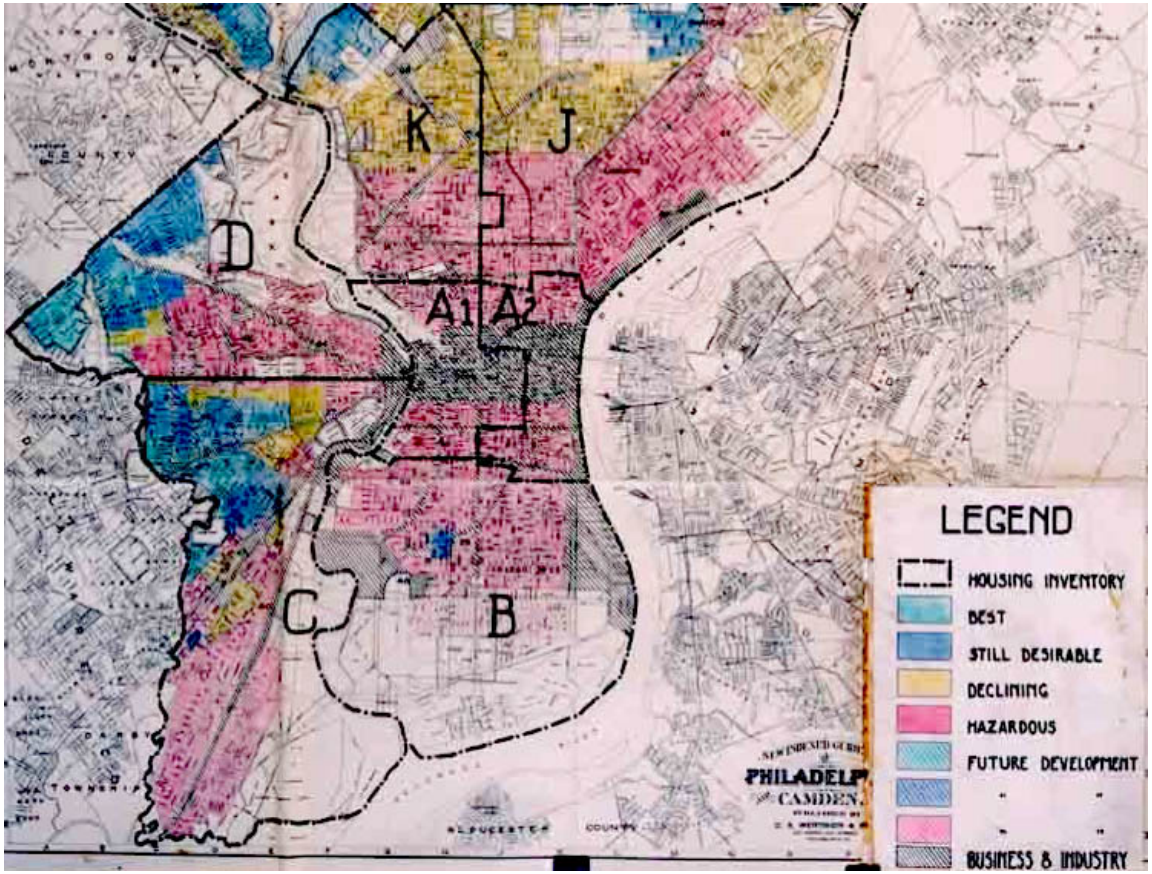
- Opacity
  - If you knew the formula, you could game it, therefore decision-makers keep their formulas secret
  - Since you don't know the formula, you don't know when it is giving undue weight to something that happened to you in an unfortunate accident
- Scale
  - Everybody wants the model that's best ***on average***
  - If everybody uses the same model, they all make the same mistake
- Damage
  - On average, a statistical model is better than a biased human
  - … but the one person for whom the model fails might have their life destroyed, especially if every decision-maker uses the same model

Examples of the problem

- **Opacity**: The “Level of Service Inventory-Revised” (LSI-R) was used to decide who gets parole in at least two states, and many counties/precincts.
  - It did not ask about race.
  - It did ask “when was your first encounter with police” and other questions that are highly correlated with race.
- **Scale**: Companies can’t use medical tests to determine hiring, but they are allowed to use personality tests. In 2016, a lawsuit found that all the employers in one metro area were using the same “personality test” to screen applicants, so people with ”undesirable” personalities could not work.
- **Damage**: The collapse of the world economy in 2008 was caused by a statistical model with a bug. Most large banks used the Gaussian copula model to decide who got home loans; it failed to correctly model the risk of multiple simultaneous defaults.

Redlining

- “Redlining” is the practice of withholding home loans or investment from people who live in “bad neighborhoods”
- Traditionally, “bad neighborhood” meant that most people who lived there were racial minorities



https://commons.wikimedia.org/wiki/File:Home_Owners%27_Loan_Corporation_Philadelphia_redlining_map.jpg

Redlining by AI

- Until recently, in many places, it was illegal for an AI to use race, gender, or ethnicity in its decision-making formula (still illegal in most of Europe)
- Many “proxy variables” correlate with race, gender, and ethnicity, e.g., home address, name, number of times you’ve had to speak to the police
- Widely-used AI decision-makers have been shown to make predictions, based on proxy variables, that are highly discriminatory in practice

Outline

- Fairness Problems
  - Opacity; Scale; Damage
  - Redlining
- Definitions of Fairness
  - Individual Fairness; Counterfactual Fairness
  - Demograpic Parity vs. Equal Odds vs. Predictive Parity
  - The 4^{th} box?

Some Published Definitions of Fairness in AI

Individual Fairness:

The dissimilarity between two outcomes should be less than the dissimilarity between the people.

$$D(f(x_1), f(x_2)) \leq d(x_1, x_2)$$

Counterfactual Fairness:

If a person's protected attribute were changed (and all their other attributes were possibly changed, according to their dependence on the protected attribute), then the outcome should not change.

Some Published Definitions of Fairness in AI

Equal Opportunity:

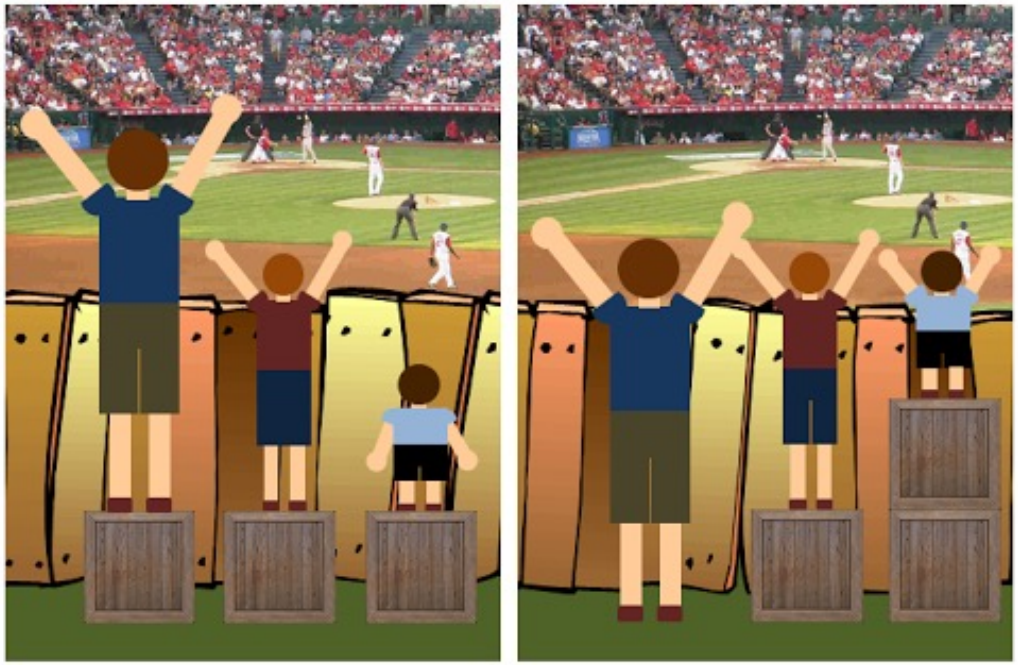
People with similar abilities succeed, regardless of irrelevant attributes.

Demographic Parity:

People succeed, regardless of irrelevant attributes.

Predictive Parity:

All successful people are qualified, regardless of irrelevant attributes.



Craig Froehle, https://medium.com/@CRA1G/the-evolution-of-an-accidental-meme-ddc4e139e0e4#.pqiclk8pl

Turning ideas into algorithms

Equal Opportunity:

People with similar abilities succeed, regardless of irrelevant attributes.

Demographic Parity:

People succeed, regardless of irrelevant attributes.

Predictive Parity:

All successful people are qualified, regardless of irrelevant attributes.

Y = a human would give you the job

$f(X)$ = AI gives you the job

A = some irrelevant attribute

Equal Opportunity:

$$\mathrm{P}(f(X)|A,Y) = \mathrm{P}(f(X)|\neg A,Y)$$

Demographic Parity:

$$\mathrm{P}(f(X)|A) = \mathrm{P}(f(X)|\neg A)$$

Predictive Parity:

$$\mathrm{P}(Y|f(X)A) = \mathrm{P}(Y|f(X),\neg A)$$

Y = a human would hire you

$f(X)$ = AI hires you

A = some irrelevant attribute

Equal Opportunity:

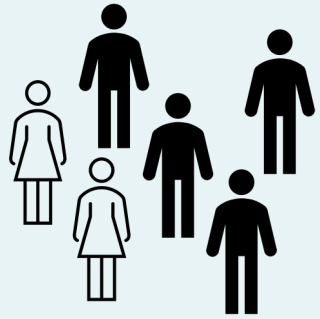
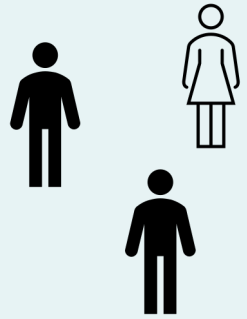
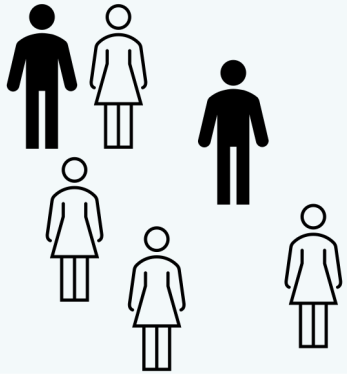
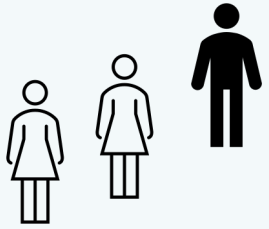
$$\mathrm{P}(f(X)|A, Y) = \mathrm{P}(f(X)|\neg A, Y)$$

Demographic Parity:

$$\mathrm{P}(f(X)|A) = \mathrm{P}(f(X)|\neg A)$$

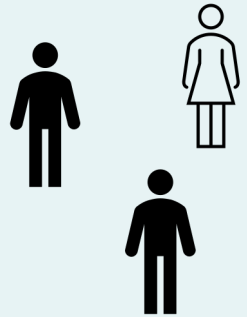
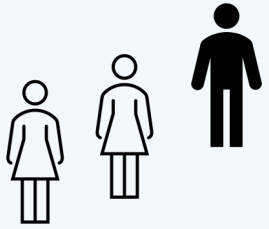
Predictive Parity:

$$\mathrm{P}(Y|f(X)A) = \mathrm{P}(Y|f(X), \neg A)$$

Confusion Matrix	$\neg f(X)$	$f(X)$
$\neg Y$		
Y		

Demographic parity: People succeed, regardless of irrelevant attributes.

$$\mathrm{P}(f(X)|A) = \mathrm{P}(f(X)|\neg A)$$

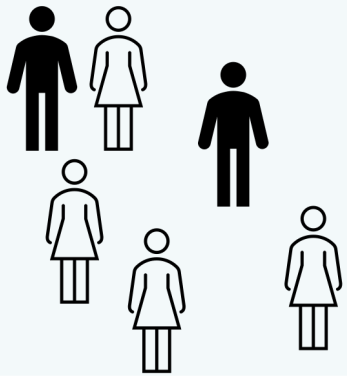
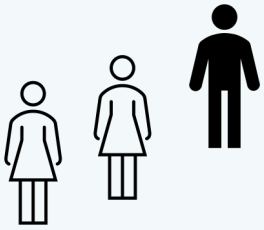
Confusion Matrix		$f(X)$
		
		

Why it matters

- Perception: This is the metric of fairness that's most visible to the public
- Revolution: If the public thinks your algorithm is unfair, they might stage massive protests to get your algorithm changed
- Generational Justice: If group X has no power, then nobody in power understands the problems that are keeping group X out of power

Equal opportunity: People with similar abilities succeed, regardless of irrelevant attributes.

$$\mathrm{P}(f(X)|A, Y) = \mathrm{P}(f(X)|\neg A, Y)$$

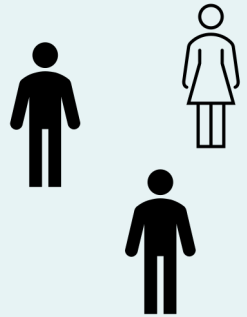
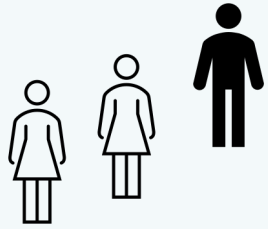
Confusion Matrix	$\neg f(X)$	$f(X)$
Y		

Why it matters

- Individual Justice: Your chance of success should only depend on your qualifications. It should not depend on irrelevant attributes

Predictive parity: All successful people are qualified, regardless of irrelevant attributes

$$\mathrm{P}(Y|f(X)A) = \mathrm{P}(Y|f(X),\neg A)?$$

Confusion Matrix		$f(X)$
$\neg Y$		
Y		

Why it matters

- Perception: If humans see those hired by the AI, and consider the hirelings from group $\neg A$ to be unqualified, they will conclude that your algorithm is unfair
- Counter-Revolution: They might conclude that science causes unfairness, and vote to cut funding for science

Your algorithm can be fair in all three ways at once only if the world is already fair

- Bayes rule:

$$P(Y|f(X),A) = \frac{P(f(X)|Y,A)P(Y|A)}{P(f(X)|A)}$$

- Demographic parity: $P(f(X)|A) = P(f(X)|\neg A)$
- Equal opportunity: $P(f(X)|Y,A) = P(f(X)|Y,\neg A)$
- Predictive parity: $P(Y|f(X),A) = P(Y|f(X),\neg A)$
- … all at once only if $P(Y|A) = P(Y|\neg A)$, i.e., humans are already hiring people from A and $\neg A$ at equal rates

Solutions?

- Sacrifice demographic parity? If $P(Y|A) > P(Y|\neg A)$, maybe the human decision-makers are right, so it's OK if $P(f(X)|A) > P(f(X)|\neg A)$?
- Sacrifice equality of opportunity? Set $P(f(X)|Y, \neg A) = 1$ but let $P(f(X)|Y, A) < 1$?
- Sacrifice predictive parity? Hire partly-qualified people from group $\neg A$, but give them on-the-job training to complete qualification?
- Change the job, so the A vs. $\neg A$ difference has less effect on people's perceptions of qualification?
- Change society? Change elementary-school education to make sure that everybody is qualified for your job?

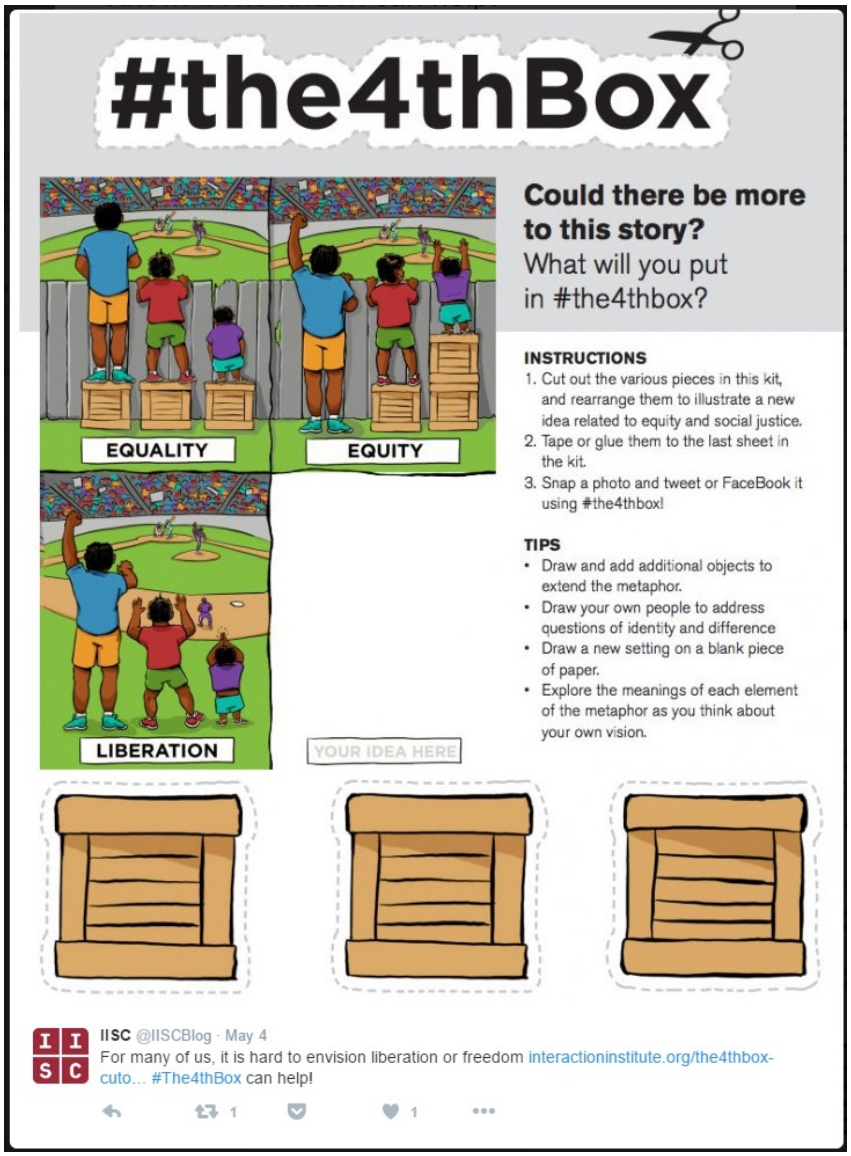
Quiz

Go to PrairieLearn, try the quiz!

Other options: Change the system?

Angus McGuire re-drew a meme originally created by Craig Froehle. And then the internet got involved…

https://medium.com/@CRA1G/the-evolution-of-an-accidental-meme-ddc4e139e0e4#.pqiclk8pl



AI: Turning ideas into algorithms

Individual Fairness:

$$D(f(x_1), f(x_2)) \leq d(x_1, x_2)$$

Equal Opportunity:

$$\mathrm{P}(f(X)|A, Y) = \mathrm{P}(f(X)|\neg A, Y)$$

Demographic Parity:

$$\mathrm{P}(f(X)|A) = \mathrm{P}(f(X)|\neg A)$$

Predictive Parity:

$$\mathrm{P}(Y|f(X), A) = \mathrm{P}(Y|f(X), \neg A)$$

#the4thBox: change the system? … how?