CS440/ECE448 Lecture 27: Fairness, Bias & Trust

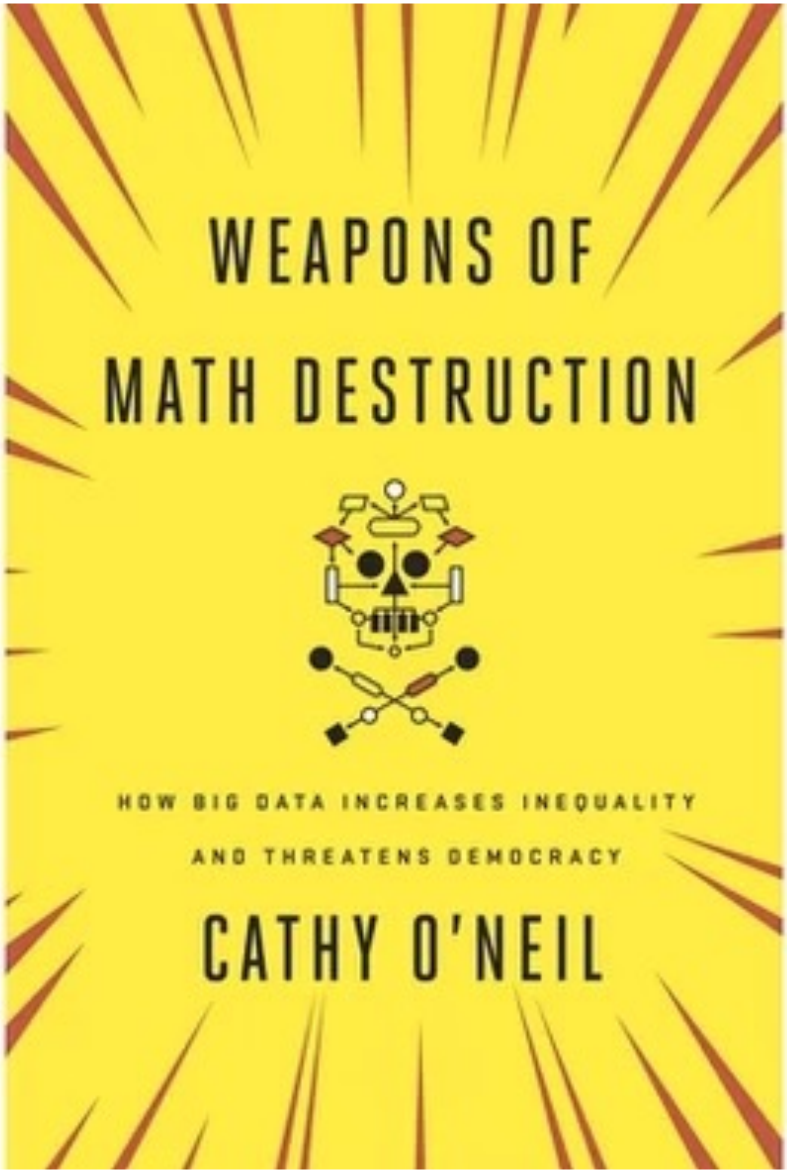
Mark Hasegawa-Johnson, 2/2022

CC-SA 2.0, Kathy Simon, 2008

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

Outline

- What’s the problem?
- Fairness
  - Representation
  - Proportion, Precision, Recall: Demographic Parity, Predictive Parity, Equalized Odds
  - Counterfactual Fairness
- Trust
  - Physical safety
  - Data safety
  - Explainable AI



WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

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**: The collapse of Lehman Brothers 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.
- **Damage**: Companies can’t use medical tests to determine hiring, but they are allowed to use personality tests. In 2016, a lawsuit found that at least seven companies were using the same personality test, and therefore rejecting the same applicants, for the same frivolous reasons.

Weapons of Math Destruction

Opacity, Scale, and Damage: a WMD is a statistical model afflicted by two of these three.

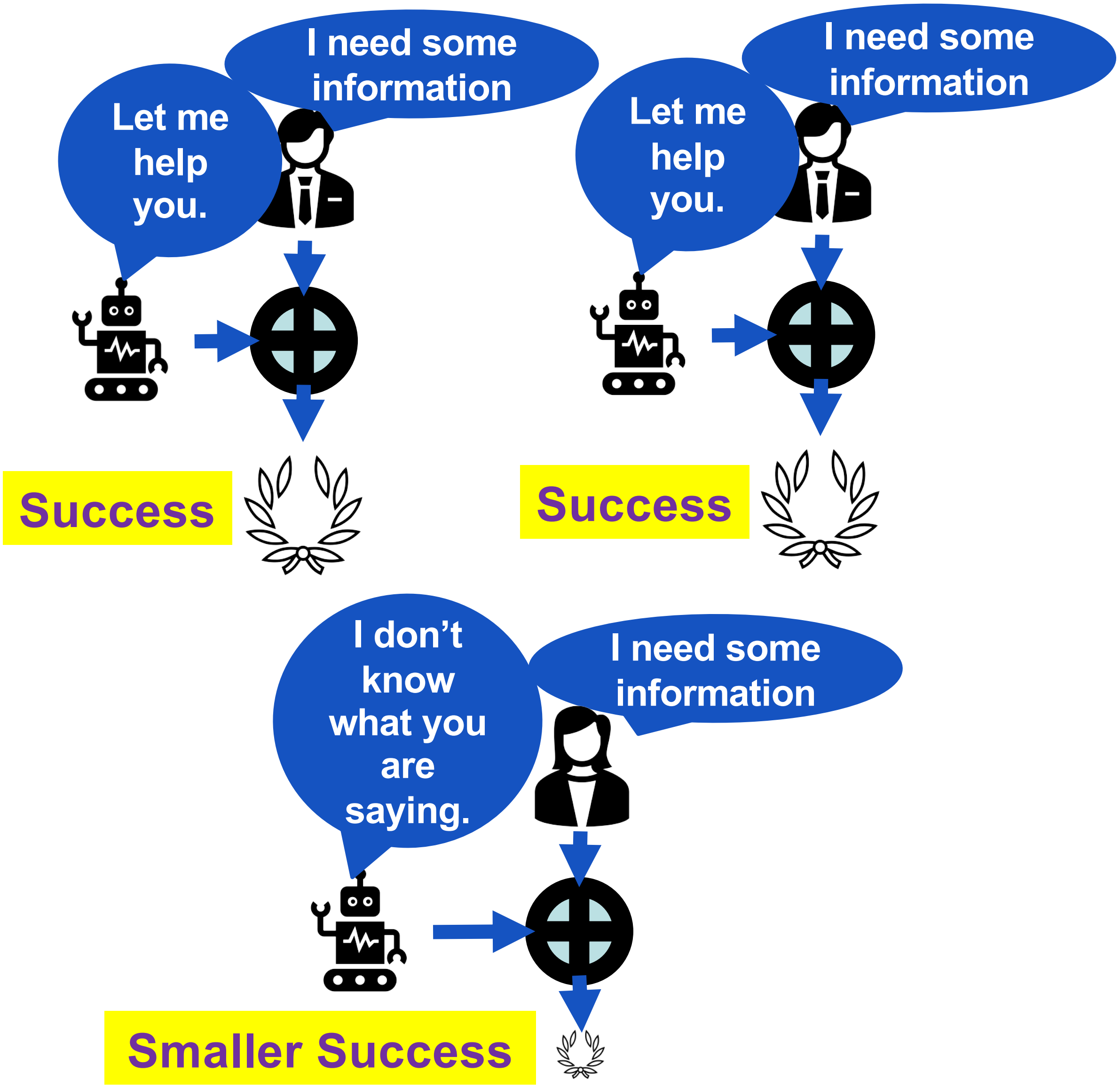
- Opacity: the relationship between inputs and outputs is hidden.
- Scale: the model is used at a scale much larger than it was ever tested for.
- Damage: negative decisions can damage people’s lives.

Developments since 2016: Scale

- UCLA had 139,500 applicants in 2021 (CBS).
- In one 24-hour period (September 16, 2020), 384,000 people applied for jobs at Amazon (Forbes).
- NeurIPS had 9454 submitted papers in 2020. They don't use AI to review the papers (yet?), but they use an automated paper-reviewer assignment system. The same system (Toronto Paper Matching System) is used by ICML, CVPR, ICCV, and ECCV.

Example: Automatic Speech Recognition

- ASR is a useful productivity tool
- If ASR works for you, it gives you a socioeconomic advantage, compared to somebody for whom ASR fails
- The people for whom ASR fails are often those who are ***already*** socioeconomically disadvantaged



Let me help you.

I need some information

Success

Let me help you.

I need some information

Success

I don't know what you are saying.

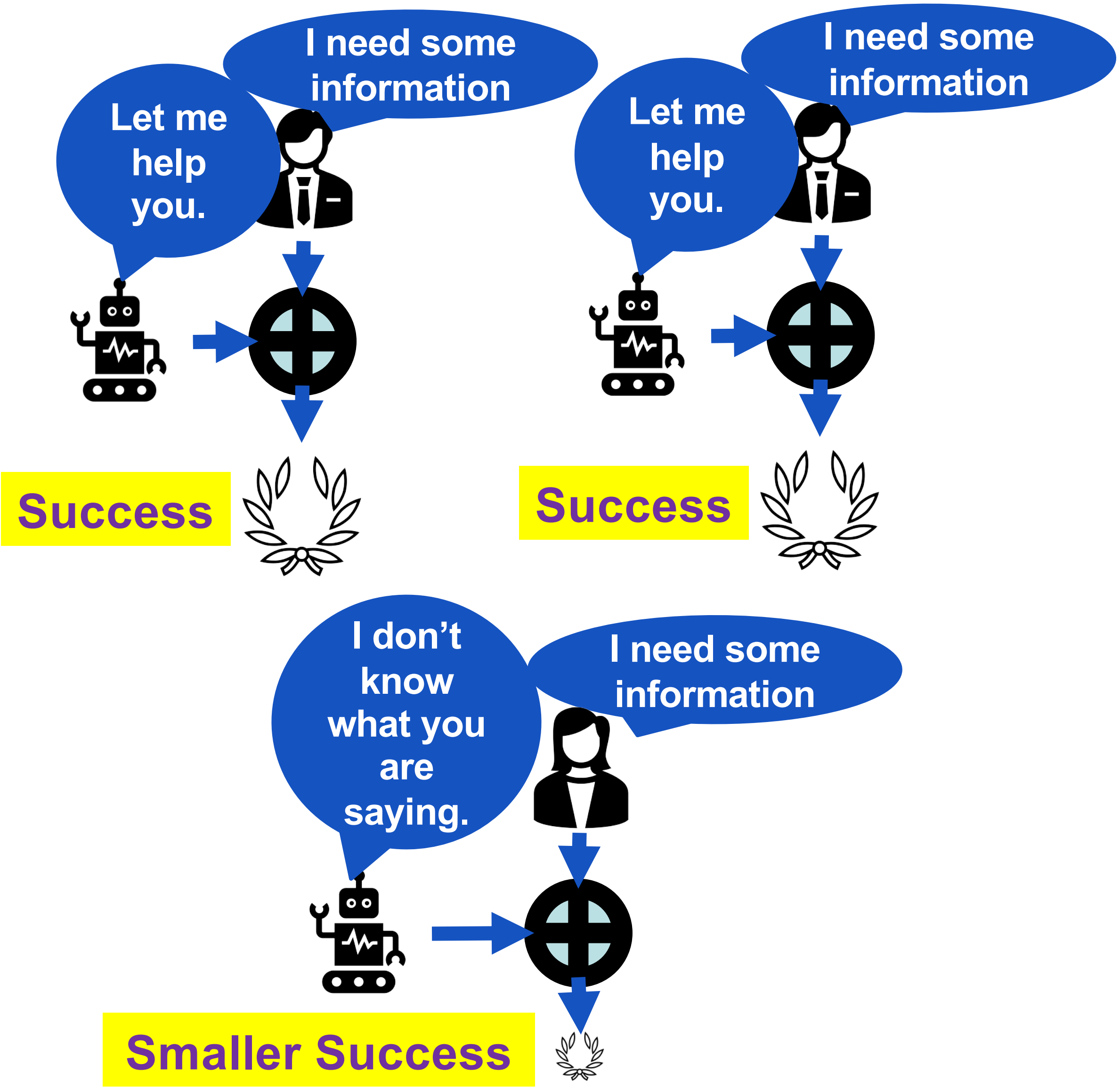
I need some information

Smaller Success

Why does ASR fail?

Reasons it may fail for a group:

- **Under-representation**: The training corpus doesn't have enough examples
- **Inter-group variance:** The group speaks differently from other groups
- **Intra-group variance:** Members of the group all speak differently from one another
- **Intra-individual variance:** Members of the group speak less precisely



Automatic Speech Recognition Word Error Rates (WER)

Gender:

- Women > Men (51% > 38%: Tatman, 2017, YouTube captioning)
- Women > Men (61% > 47%: Garnerin et al., 2019, European broadcast news)
- Black Men > Black Women (41% > 30%: Koenecke et al., 2020)

Dialect:

- Scottish > American (53% > 42%: Tatman, 2017)
- American Deep South > General American (Picone 1991)

Race:

- Black > White (35% > 19%: Koenecke: avg of Amazon, Apple, Google, IBM, Microsoft)

Disability:

- People w/Cerebral Palsy > People w/o (41% > 33%: Issa et al., in review)

Age:

- Teenage (<20)) > Old (>70) > Young adult (20-30) > Old adult (50-70) (Feng & Scharenborg, 2020, Sarı et al., 2021)

Outline

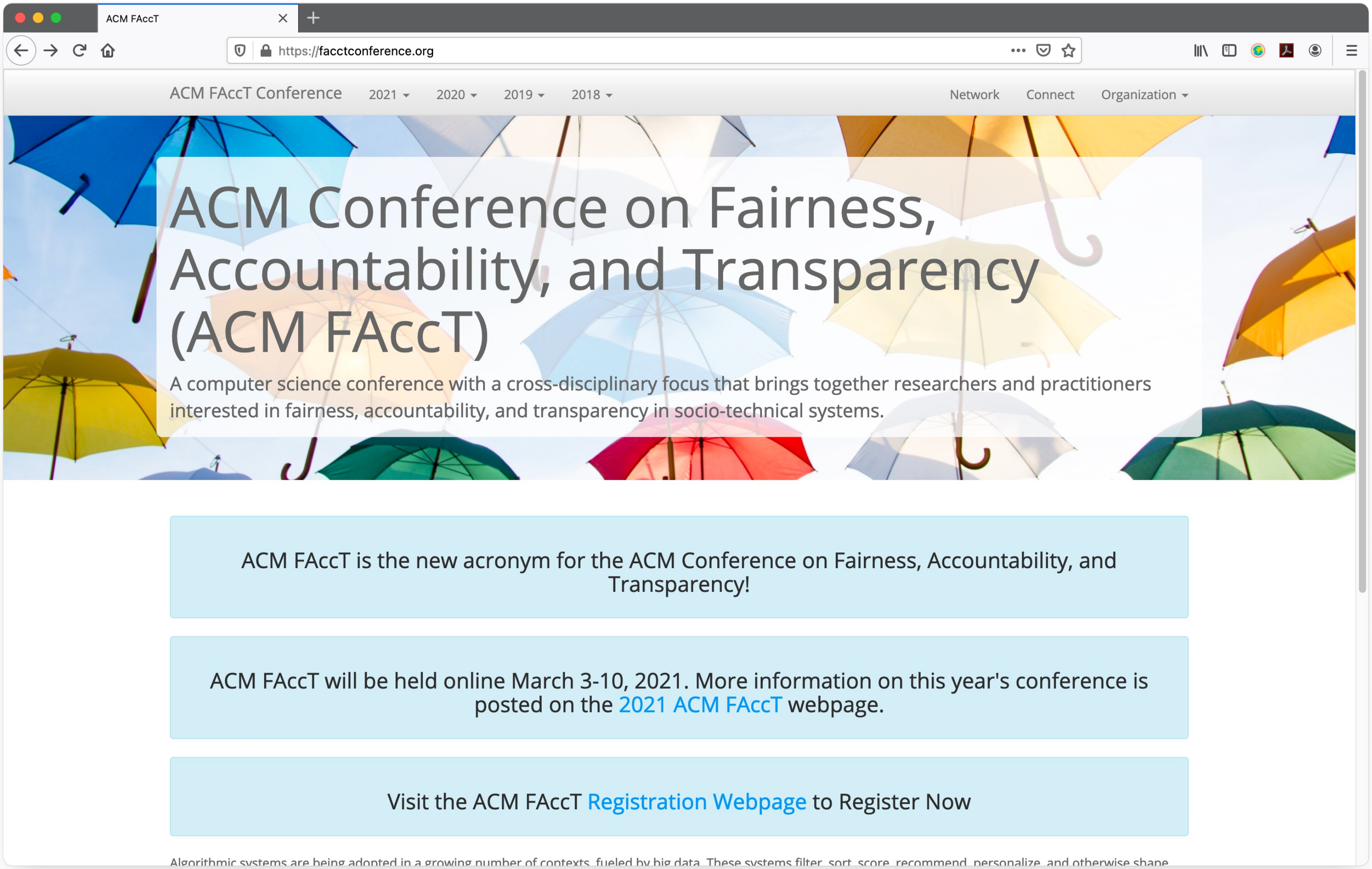
- What's the problem?
- Fairness
  - Representation
  - Proportion, Precision, Recall: Demographic Parity, Predictive Parity, Equalized Odds
  - Counterfactual Fairness
- Trust
  - Physical safety
  - Data safety
  - Explainable AI

Bias caused by Data Sparsity

- Data contain more examples of one type than others, e.g., more Caucasians than African Americans
- Accuracy may be higher for the type that is better represented in the training data (minimize error by minimizing error for the majority case)
- Example: blacks more likely to be refused parole even if their prison records are the same (https://www.nytimes.com/2016/12/04/nyregion/new-york-prisons-inmates-parole-race.html)
- Example: tweets containing African American vernacular classified as “Danish,” and therefore excluded from automatic sentiment analysis (https://www.technologyreview.com/s/608619/ai-programs-are-learning-to-exclude-some-african-american-voices/)

Some possible answers

- Governments and private organizations now have funded efforts to acquire more data from under-represented groups.
  - Corpus of Regional African-American Language
  - Bureau of Justice Statistics
  - NIH Inclusion Policies for Research Involving Human Subjects
- Academia and industry seek to increase representation in AI data by increasing diversity among AI experts
  - AI4ALL



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.

ACM FAccT is the new acronym for the ACM Conference on Fairness, Accountability, and Transparency!

ACM FAccT will be held online March 3-10, 2021. More information on this year's conference is posted on the 2021 ACM FAccT webpage.

Visit the ACM FAccT Registration Webpage to Register Now

Algorithmic systems are being adopted in a growing number of contexts, fueled by big data. These systems filter, sort, score, recommend, personalize, and otherwise shape

Automatic Speech Recognition: Database Collection Efforts

Targeted to particular user groups:

- UASpeech: dysarthria as a symptom of Cerebral Palsy
  - 3.2 hours
- CORAAL (Corpus of Regional African American Languages)
  - 200 hours

Recent efforts with lots of speech from lots of people; we currently have no idea what is the distribution across race, ethnicity, gender, orientation, native language, etc., but we're working on it:

- 100,000 Podcasts Corpus
  - 40,000 hours
- The People's Speech (recordings of town hall meetings, etc)
  - 30,000 hours

Outline

- What's the problem?
- Fairness
  - Representation
  - Proportion, Precision, Recall: Demographic Parity, Predictive Parity, Equalized Odds
  - Counterfactual Fairness
- Trust
  - Physical safety
  - Data safety
  - Explainable AI

Standard Definitions of Fairness in AI

Let’s define the following random variables:

- A = protected attribute: An observable fact that should not be predictive of outcomes, e.g., gender, race, age, disability.
- X = observable data that we can use for our decision
- Y = the unknown correct label for this person (e.g., $Y = 1$ might mean “this person should receive a loan” or “should be admitted to UIUC”)
- $f(X)$ = a function of X, designed using probabilistic or neural methods to approximate Y as closely as possible

Standard Definitions of Fairness in AI

Proportion a.k.a. **Demographic Parity:**

The probability of a positive outcome is the same, regardless of protected attribute.

$$P(f(X) = 1|A = a) = P(f(X) = 1|A = a') \quad \forall a, a'$$

Precision a.k.a. **Predictive Parity:**

Precision is the same, regardless of protected attribute.

$$P(Y = 1|f(X) = 1, A = a) = P(Y = 1|f(X) = 1, A = a') \quad \forall x, a, a'$$

Recall a.k.a. **Equalized Odds:**

Recall is the same, regardless of protected attribute.

$$P(f(X) = 1|Y = 1, A = a) = P(f(X) = 1|Y = 1, A = a') \quad \forall x, a, a'$$

You can’t have all three

$$P(f(X) = 1|Y = 1, A = a) = \frac{P(Y = 1|f(X) = 1, A = a)P(f(X) = 1|A = a)}{P(Y = 1|A = a)}$$

$$P(f(X) = 1|Y = 1, A = a') = \frac{P(Y = 1|f(X) = 1, A = a')P(f(X) = 1|A = a')}{P(Y = 1|A = a')}$$

The balanced error, predictive parity, and demographic parity terms cannot all be independent of A unless Y is also independent of A.

In other words, if the current state of society is unfair (distribution of positive outcomes currently depends on protected attribute), then algorithmic solutions cannot make it fair (at least not in all three ways, all at once).

Other problems with algorithmic solutions

Dwork (2012) points out that demographic parity can lead to socially undesirable outcomes, e.g., people gaming the system.

… but …

Srivastava, Heidari and Krause (2019) found that users of an AI judge its fairness based on demographic parity. They ignore predictive parity and balanced error, even when these concepts are explained to them.

Other Useful Definitions of Fairness in AI

Individual Fairness:

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

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.

Outline

- What's the problem?
- Fairness
  - Representation
  - Proportion, Precision, Recall: Demographic Parity, Predictive Parity, Equalized Odds
  - Counterfactual Fairness
- Trust
  - Physical safety
  - Data safety
  - Explainable AI

Are College Admissions Fair?

- Bickel, Hammel, and O’Connell, “Sex bias in graduate admissions: Data from Berkeley,” Science 187(4175):398–404, 1975
- At that time, women were being admitted to Berkeley at a far lower rate than men:

$$P(f(X) = \mathrm{admit} | A = \mathrm{female}) < P(f(X) = \mathrm{admit} | A = \mathrm{male})$$

Are College Admissions Fair?

- Bickel, Hammel, and O’Connell added one more variable to the analysis: Z=Department to which the student applied
- They found that, for each individual department, men and women were being admitted with equal probability

$$P(f(X) = \text{admit} | Z, A = \text{female}) = P(f(X) = \text{admit} | Z, A = \text{male})$$

- Question to ponder: does this make the outcome fair? BHO said “yes,” but other people said “no.” Debate still rages.

Analyzing Fairness Using Bayesian Networks

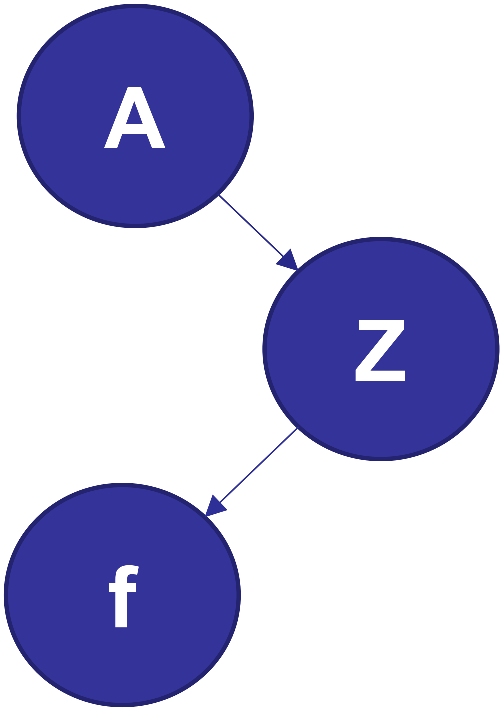
Pearl analyzed this result using the Bayesian network at right. It fits the data with four sets of parameters:

$$P(A = \text{female}) \approx \frac{1}{2}$$

$$P(Z|A = \text{female})$$

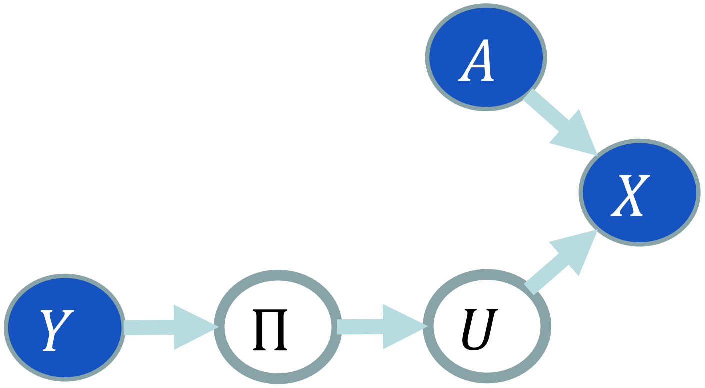
$$P(Z|A = \text{male})$$

$$P(f(X) = \text{admit}|Z)$$



Counterfactually Fair Automatic Speech Recognition

Sarı, Hasegawa-Johnson & Yoo, 2021



- **Open circles denote unobserved variables**
- **Filled circles denote variables that are observed during training**

Some assumptions:

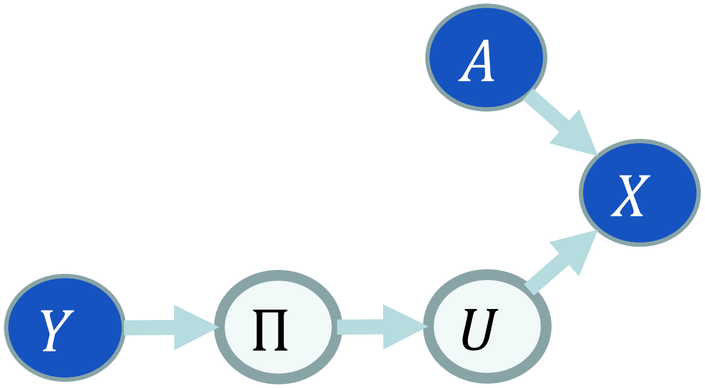
- A $=$ **protected attribute (gender, race, dialect, age, education, disability)**
- Y $=$ **text of the words that are spoken (**$\perp A$**)**
- X $=$ **person-dependent acoustic signal**

Some things that we design, so that they are independent of A:

- $\Pi =$ **time alignment of text to audio (**$\perp A$**)**
- U $=$ **hidden layer activations of the neural net (**$\perp A$**)**

Counterfactually Fair Automatic Speech Recognition

Sarı, Hasegawa-Johnson & Yoo, in review



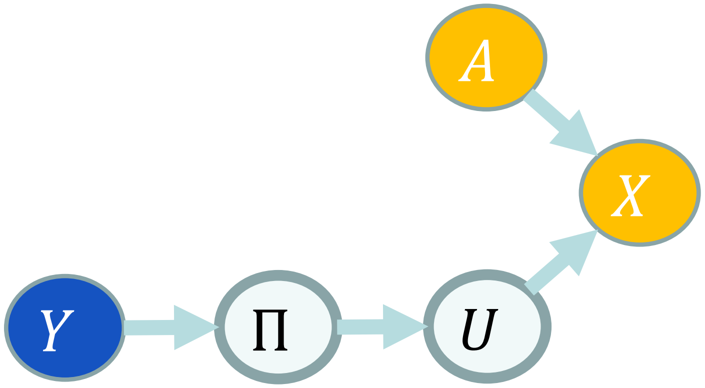
Lightly-shaded circles denote variables whose values are inferred based on models of joint probability distribution

Abduction:

- **Infer U and Π from X, Y and A**

Counterfactually Fair Automatic Speech Recognition

Sarı, Hasegawa-Johnson & Yoo, in review



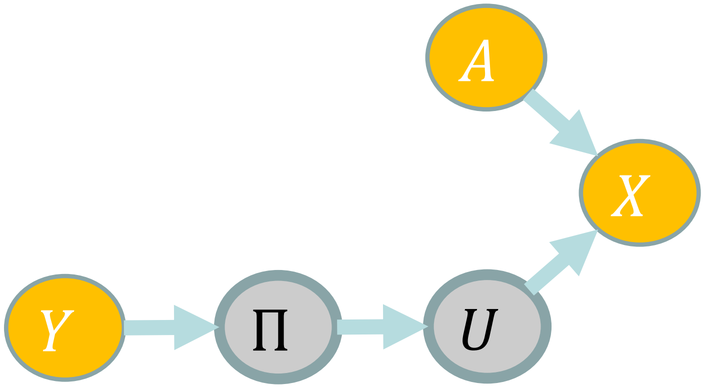
Change in color denotes counterfactual action upon a variable

Action:

- **Change A (e.g., male→female, old → young, etc.).**
- **Change X, using the new value of A, and the previously inferred value of U.**

Counterfactually Fair Automatic Speech Recognition

Sarı, Hasegawa-Johnson & Yoo, in review

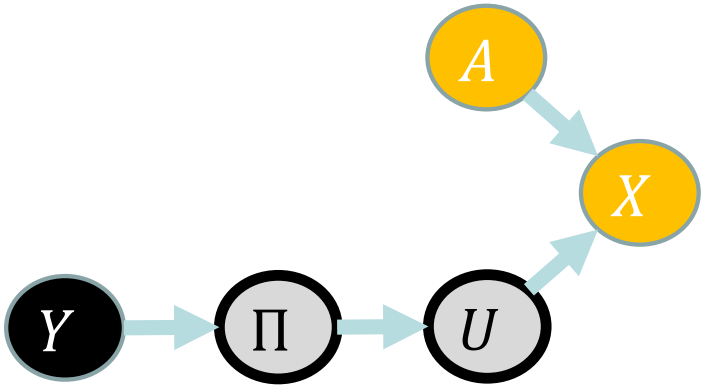


Prediction:

- **Re-compute the values of U, Π, and Y from the modified versions of X and A**

Counterfactually Fair Automatic Speech Recognition

Sarı, Hasegawa-Johnson & Yoo, in review

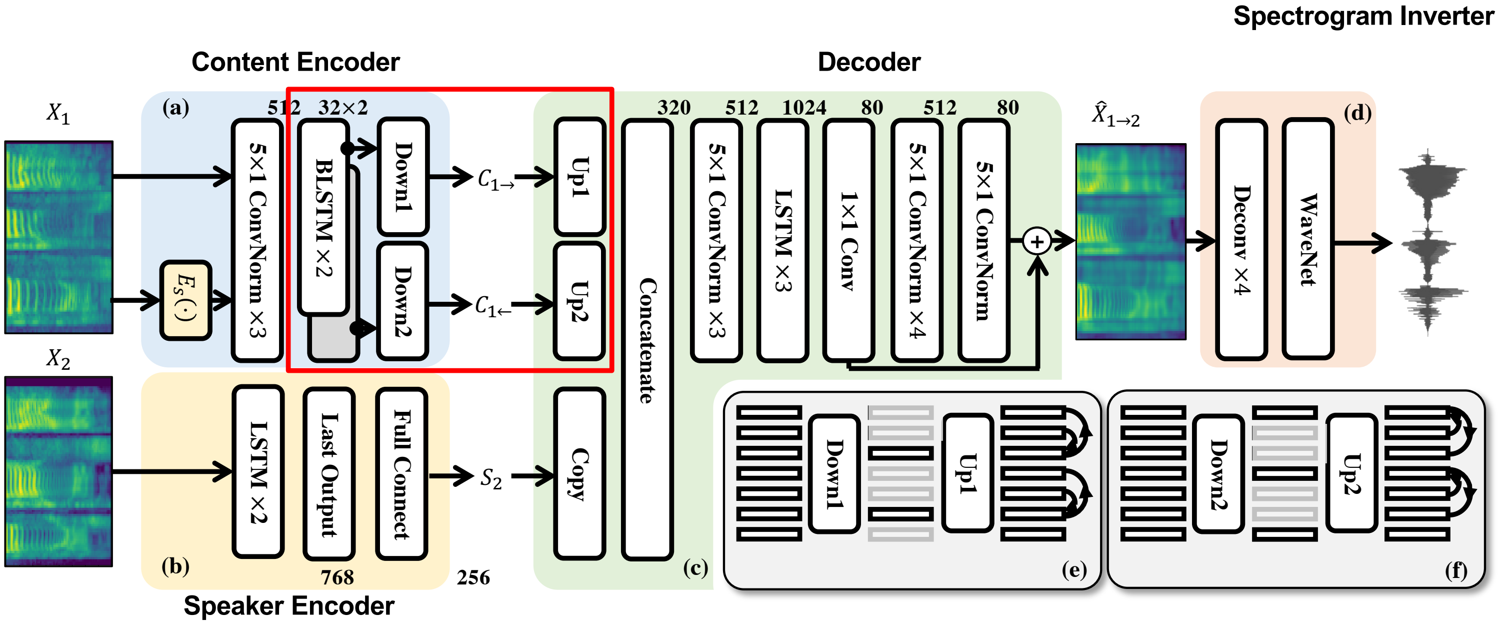


Regularization:

- **Train the neural net so that, no matter what the value of A is, U, Π are unchanged**

Action: change male->female, old-> young … How?
By using a voice conversion neural net.

AutoVC: Zero-shot voice style transfer with only autoencoder loss (Qian et al., 2019)



AutoVC: Zero-shot voice style transfer with only autoencoder loss

Qian, Zhang, Chang, Yang, and Hasegawa-Johnson, 2019

- To hear is to believe
- Here are the reference speakers, in their own voices:

	T1 🔊	T2 🔊	T3 🔊
Sa	🔊	🔊	🔊
Sb	🔊	🔊	🔊

- Adversarial game: Utterance will be “Please call Stella.” All three (T1, T2, and T3) are resynthesized from the content code of one (T1, T2, or T3). Which one?
- Who is Sa? T1, T2 or T3?
- Who is Sb? T1, T2 or T3?

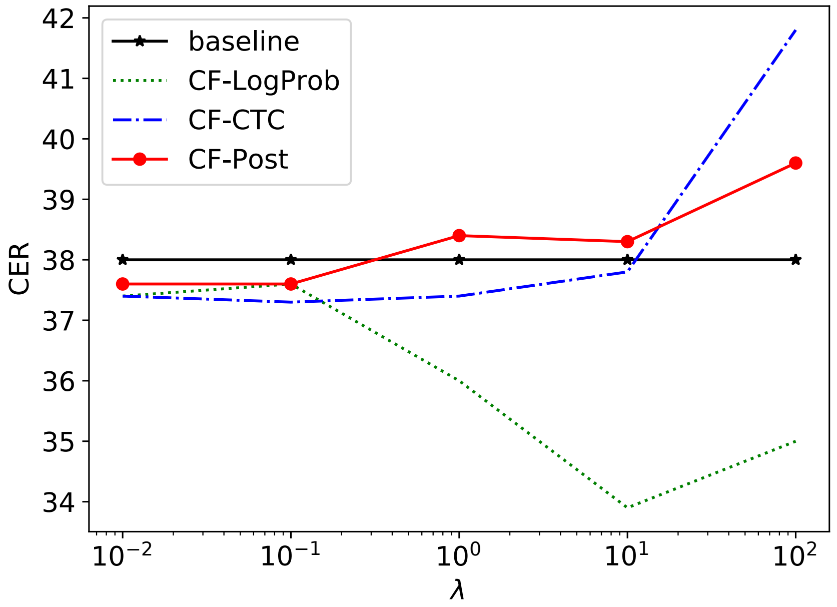
Average error rate, averaged across all speakers, as a function of regularization weight

Train the neural net using regularized training so that…

- CTC = P(sentence|audio)…
- Post = P(character|spectrum) …
- LogProb = log P(character|spectrum)…

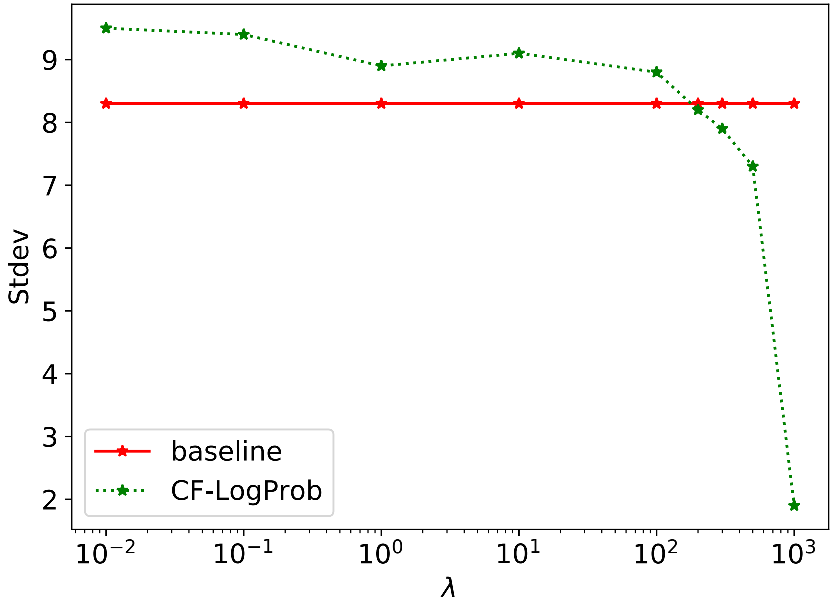
... is independent of factual vs. counterfactual group identity

CER (character error rate) = a measure of average performance across all groups



Standard deviation of error rate across speakers, as a function of regularization weight

Training criterion = speech recognizer loss + $\lambda\times$(counterfactual matching loss)



Stdev (of error rate, across speakers) = a measure of unfairness of the ASR

Model	Avg	Stdev	GAE	AAL	Diff
Two different systems	24.8	15.6	9.8	38.0	28.2
Use GAE system for both	29.2	20.1	9.8	46.4	36.6
Counterfactual data, w/o counterfactual regularization ($\lambda = 0$)	23.7	11.6	14.6	34.9	20.3
Counterfactual data & counterfactual regularization ($\lambda = 10$)	21.1	10.8	12.6	31.5	18.9
Counterfactual data & counterfactual regularization ($\lambda = 50$)	23.3	10.6	15.3	33.4	18.2

Character Error Rates (lower is better):

- Avg = Average across all speakers
- Stdev = Standard deviation among speakers
- GAE = General American English (Librispeech corpus)
- AAL = African American Language (CORAAL corpus)
- Diff = $\mathrm{AAL} - \mathrm{GAE}$

Model	Avg	Stdev	GAE	AAL	Diff
Two different systems	24.8	15.6	9.8	38.0	28.2
Use GAE system for both	29.2	20.1	9.8	46.4	36.6
Counterfactual data, w/o counterfactual regularization ($\lambda = 0$)	23.7	11.6	14.6	34.9	20.3
Counterfactual data & counterfactual regularization ($\lambda = 10$)	21.1	10.8	12.6	31.5	18.9
Counterfactual data & counterfactual regularization ($\lambda = 50$)	23.3	10.6	15.3	33.4	18.2

Observations:

1. You can’t just use the GAE system to transcribe AAL (46.4% error)

Model	Avg	Stdev	GAE	AAL	Diff
Two different systems	24.8	15.6	9.8	38.0	28.2
Use GAE system for both	29.2	20.1	9.8	46.4	36.6
Counterfactual data, w/o counterfactual regularization ($\lambda = 0$)	23.7	11.6	14.6	34.9	20.3
Counterfactual data & counterfactual regularization ($\lambda = 10$)	21.1	10.8	12.6	31.5	18.9
Counterfactual data & counterfactual regularization ($\lambda = 50$)	23.3	10.6	15.3	33.4	18.2

Observations:

1. You can’t just use the GAE system to transcribe AAL (46.4% error)
2. Generating synthetic training data helps AAL, but harms GAE

Model	Avg	Stdev	GAE	AAL	Diff
Two different systems	24.8	15.6	9.8	38.0	28.2
Use GAE system for both	29.2	20.1	9.8	46.4	36.6
Counterfactual data, w/o counterfactual regularization ($\lambda = 0$)	23.7	11.6	14.6	34.9	20.3
Counterfactual data & counterfactual regularization ($\lambda = 10$)	21.1	10.8	12.6	31.5	18.9
Counterfactual data & counterfactual regularization ($\lambda = 50$)	23.3	10.6	15.3	33.4	18.2

Observations:

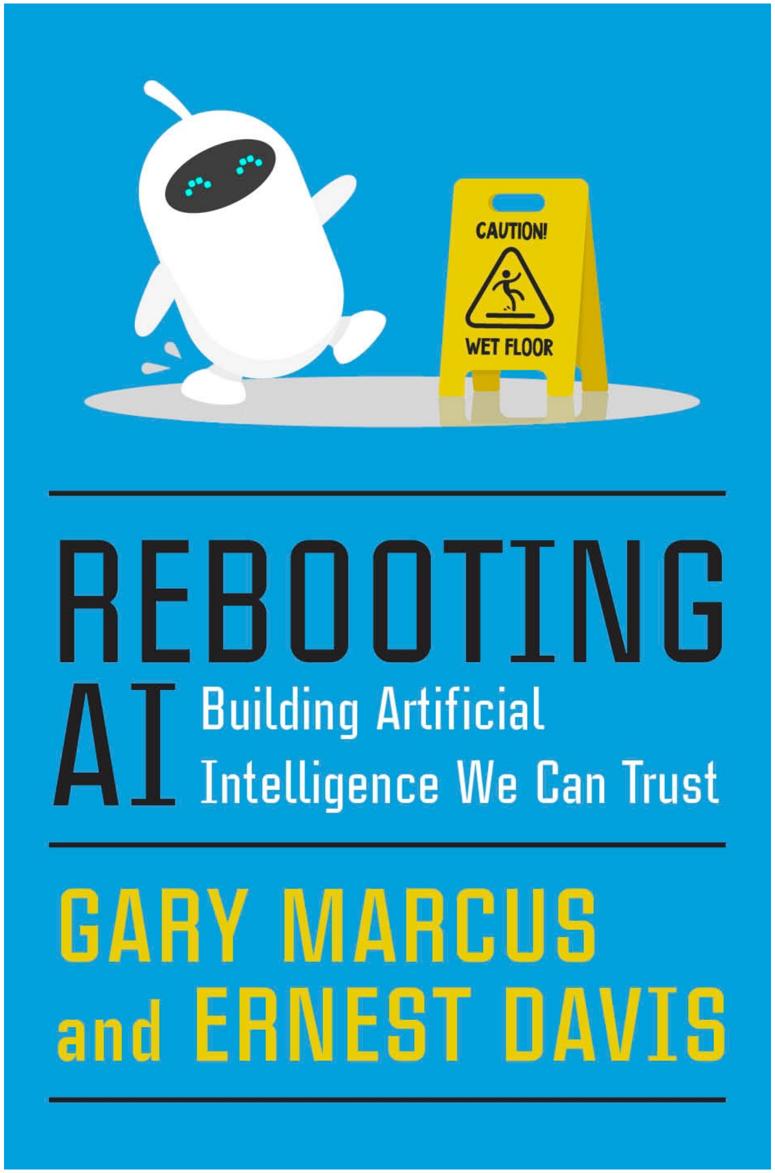
1. You can’t just use the GAE system to transcribe AAL (46.4% error)
2. Generating synthetic training data helps AAL, but harms GAE
3. Counterfactual regularization helps (less harm to GAE, more benefit to AAL).

Notes

- What we have done:
  - Used GAE data to lower AAL error rates
  - … at the expense of higher GAE error rates.
- Is that desirable? It depends on your customers:
  - Minimin use case: Minimize error for the low-error users. Sell your product to only the people for whom it works.
  - Minimax use case: Minimize error for the high-error users. Sell your product to everybody.

Outline

- What’s the problem?
- Fairness
  - Representation
  - Proportion, Precision, Recall: Demographic Parity, Predictive Parity, Equalized Odds
  - Counterfactual Fairness
- Trust
  - Physical safety
  - Data safety
  - Explainable AI



REBOOTING AI

Building Artificial Intelligence We Can Trust

GARY MARCUS and ERNEST DAVIS

Trust Issues

- Physical Safety
  - April 18, 2021: 2 Killed in Driverless Tesla Car Crash
- Data Safety
  - March 2020: CAM4 data breach exposed 10 billion records

Physical Safety

- Robustness to changes in data distribution
- Avoiding catastrophic “edge cases”
- Robustness to adversarial examples or attacks
- Avoiding negative side effects in reward function
- Avoiding “reward hacking”

- Reading: Concrete AI safety problems

The Virtual Sully Research Project

Creative commons 2.0, multichill, 2009

- At 3:27 on 1/15/2009, US Airways 1549 lost power in both engines.
- Capt. “Sully” Sullenberger tried to turn back to LaGuardia, then tried to turn toward Teterboro, then realized there was no time.
- At 3:31 he landed the plane in the Hudson river.
- All passengers were saved.

The Virtual Sully Research Project

Creative commons 2.0, multichill, 2009

Virtual Sully research project seeks to give AI

- the ability to plan a course of action with backup plans available in case of unexpected disaster,
- the ability to quickly discard low-priority goals in favor of threatened high-priority goals in case of the unexpected inability to achieve both.

Outline

- What's the problem?
- Fairness
  - Representation
  - Proportion, Precision, Recall: Demographic Parity, Predictive Parity, Equalized Odds
  - Counterfactual Fairness
- Trust
  - Physical safety
  - Data safety
  - Explainable AI

Data Safety

"Passports, however, use a different technology known as RFID (or Radio Frequency Identification), the same type used to tag clothing, pets, even artificial replacements for hips and knees. When embedded in a U.S. passport, the chip can be scanned only by someone at close range with an RFID reader, usually within a couple feet…

"Yes, someone nearby could read what's in your wallet. That's why I keep my passport in an RFID-shielded wallet," said G. Mark Hardy, president of National Security Corp., based in Rosedale, Md., which provides cybersecurity expertise to government and corporate clients.

But, he said, "it's less likely to happen, at this point in time, because it's so much easier to do fraud some other way.""

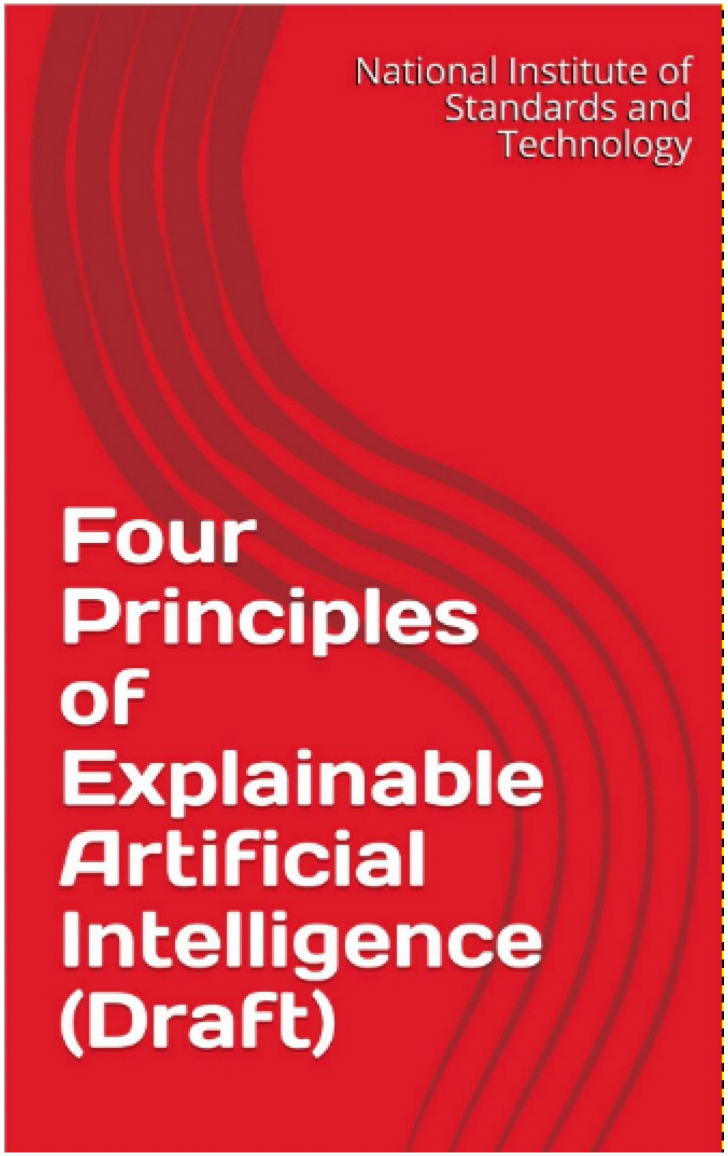
Read more here: http://www.sacbee.com/news/business/personal-finance/claudia-buck/article2599038.html#storylink=cpy

Example of a Technical Solution: Homomorphic Encryption

1. Encrypt the data on your cell phone
2. Send the encrypted data to a server
3. The server sends it through a neural net in its encrypted form, without ever decrypting it
4. They send you the result, and you decrypt it using the same key

Outline

- What's the problem?
- Fairness
  - Representation
  - Proportion, Precision, Recall: Demographic Parity, Predictive Parity, Equalized Odds
  - Counterfactual Fairness
- Trust
  - Physical safety
  - Data safety
  - Explainable AI



National Institute of
Standards and
Technology

Four
Principles
of
Explainable
Artificial
Intelligence
(Draft)

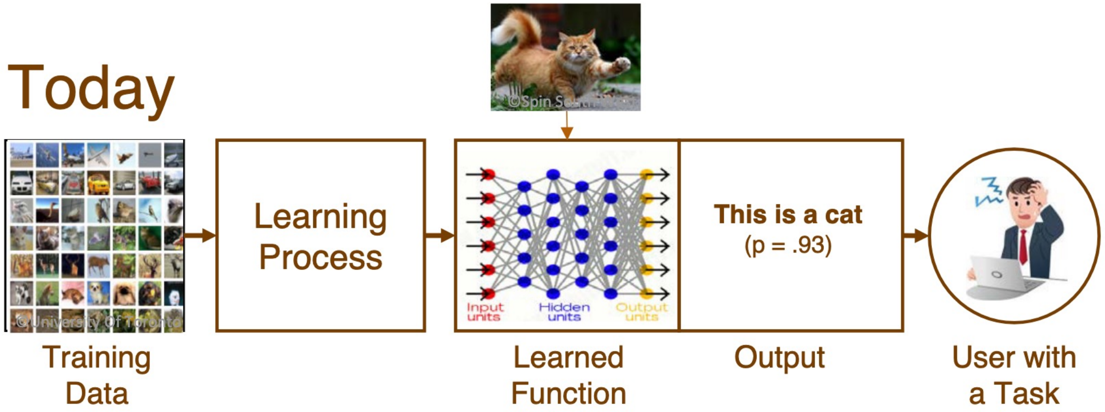
Four Principles of Explainable AI

- Explanation: The system can explain its reasons for any decision
- Meaningful: The explanation can be understood by the user
- Explanation Accuracy: The explanation correctly describes how the system made its decision
- Knowledge Limits: The system is only used under circumstances for which it was designed.

Explainable AI – What Are We Trying To Do?

David Gunning, “Explainable Artifcial Intelliigence (XAI),” 2017

Today



Training Data

Learning Process

Learned Function

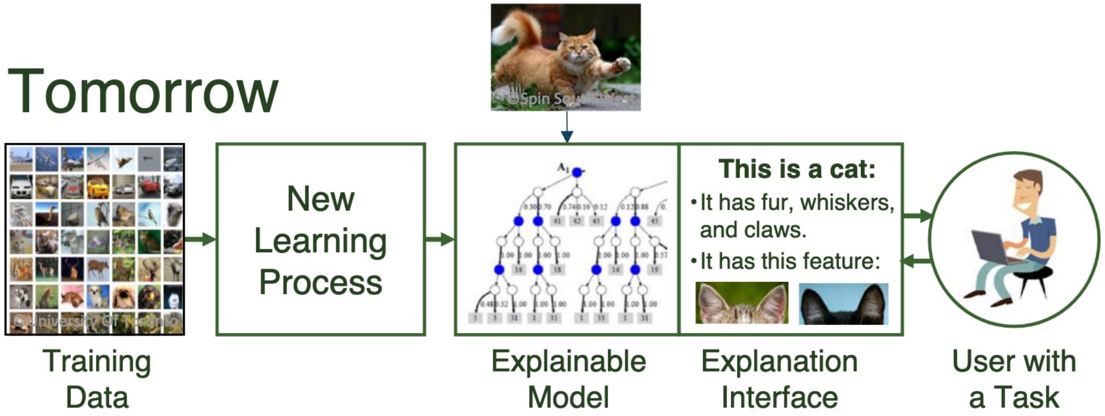
This is a cat (p = .93)

Output

User with a Task

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow



Training Data

New Learning Process

Explainable Model

This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:

Explanation Interface

User with a Task

- I understand why
- I understand why not
- I know when you’ll succeed
- I know when you’ll fail
- I know when to trust you
- I know why you erred

Methods

- Visualization
  - Pro: provides intuitively useful descriptions of typical behavior
  - Con: post-hoc explanation of the typical behavior; may not tell you much about worst-case behavior
- Causal Graphs/Bayesian Networks
  - Pro: describes reasoning process of the AI exactly
  - Con: constraining AI reasoning process to obey an explainable causal graph sometimes harms accuracy

Outline

- What's the problem?
- Fairness
  - Representation
  - Proportion, Precision, Recall: Demographic Parity, Predictive Parity, Equalized Odds

$$P(f(X) = 1|A = a) = P(f(X) = 1|A = a')$$
$$P(Y = 1|f(X) = 1, A = a) = P(Y = 1|f(X) = 1, A = a')$$
$$P(f(X) = 1|Y = 1, A = a) = P(f(X) = 1|Y = 1, A = a')$$

  - Counterfactual Fairness
- Trust
  - Physical safety
  - Data safety
  - Explainable AI