Hearing #7 on Competition and Consumer Protection in the 21st Century

Howard University

School of Law

November 14, 2018

Welcome

We Will Be Starting Shortly

Welcome and Introductory Remarks

Bruce Hoffman

Federal Trade Commission

Bureau of Competition

Algorithmic Collusion

Session moderated by:

Ellen Connelly
Federal Trade Commission
Office of Policy Planning

James Rhilinger
Federal Trade Commission
Bureau of Competition

Algorithmic Collusion

Maurice E. Stucke

University of Tennessee College of Law

Algorithmic Collusion

Ai Deng

Bates White

Algorithmic Collusion

Kai-Uwe Kühn

University of East Anglia

Algorithmic Collusion

Rosa M. Abrantes-Metz

Global Economics Group

Algorithmic Collusion

Sonia Kuester Pfaffenroth

Arnold & Porter

Algorithmic Collusion

Joseph E. Harrington, Jr.
University of Pennsylvania

Algorithmic Collusion

Panel Discussion:

Maurice E. Stucke, Ai Deng, Kai-Uwe Kühn,
Rosa M. Abrantes-Metz,
Sonia Kuester Pfaffenroth,
Joseph E. Harrington, Jr.,

Moderators: Ellen Connelly & James Rhilinger

Break

10:45-11:00 am

Framing Presentation (prerecorded)

Michael I. Jordan

University of California, Berkeley

Emerging Competition, Innovation, and Market Structure Questions Around Algorithms, Artificial Intelligence, and Predictive Analytics

Session moderated by:

Brian O'Dea
Federal Trade Commission
Bureau of Competition

Nathan Wilson
Federal Trade Commission
Bureau of Economics

Emerging Competition, Innovation, and Market Structure Questions Around Algorithms, Artificial Intelligence, and Predictive Analytics

Panel Discussion:

Robin Feldman, Joshua Gans,
Preston McAfee, Nicolas Petit

Moderators: Brian O'Dea & Nathan Wilson

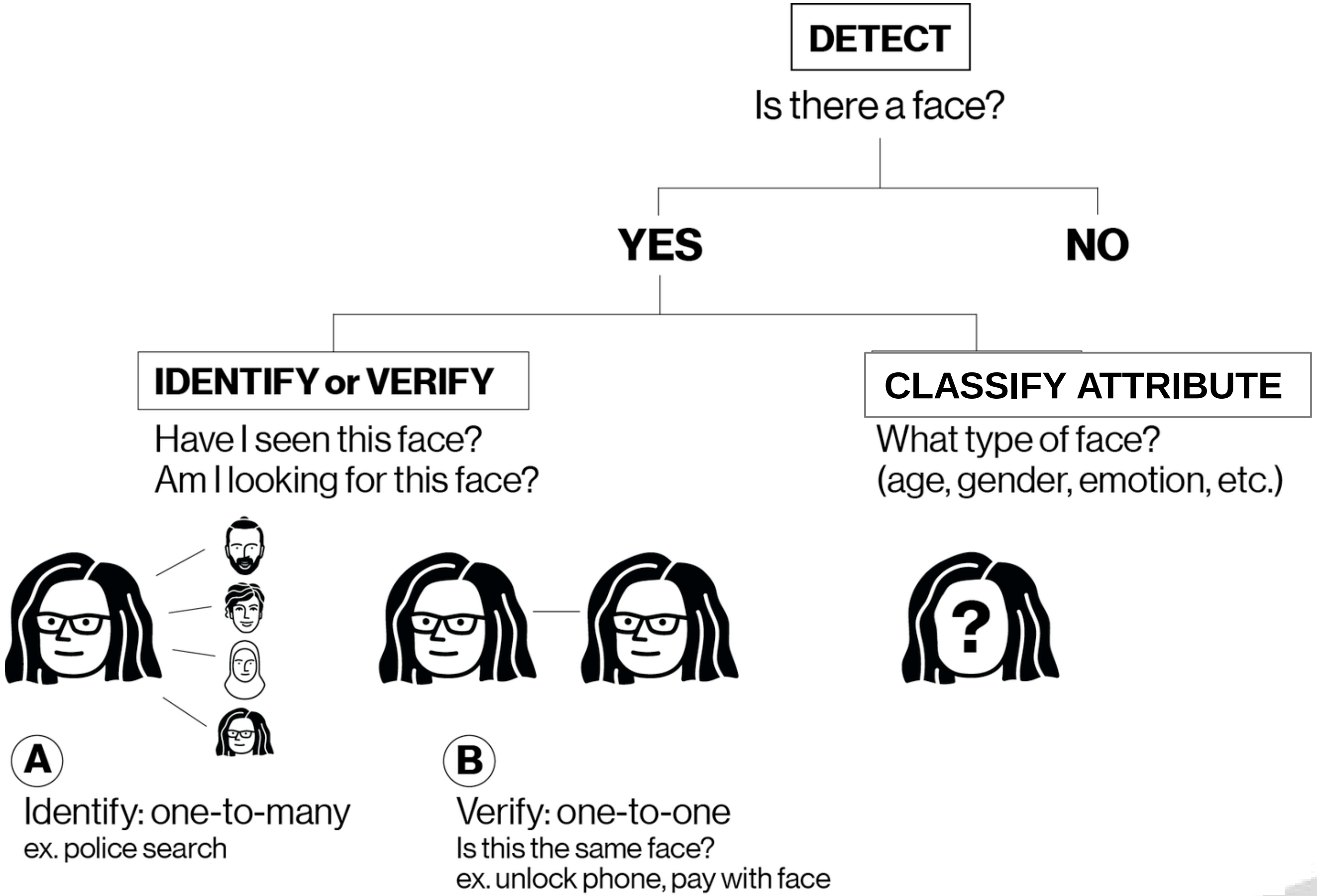
Facial Analysis Technology Warning Signs

Joy Buolamwini

Algorithmic Justice League | MIT Media Lab

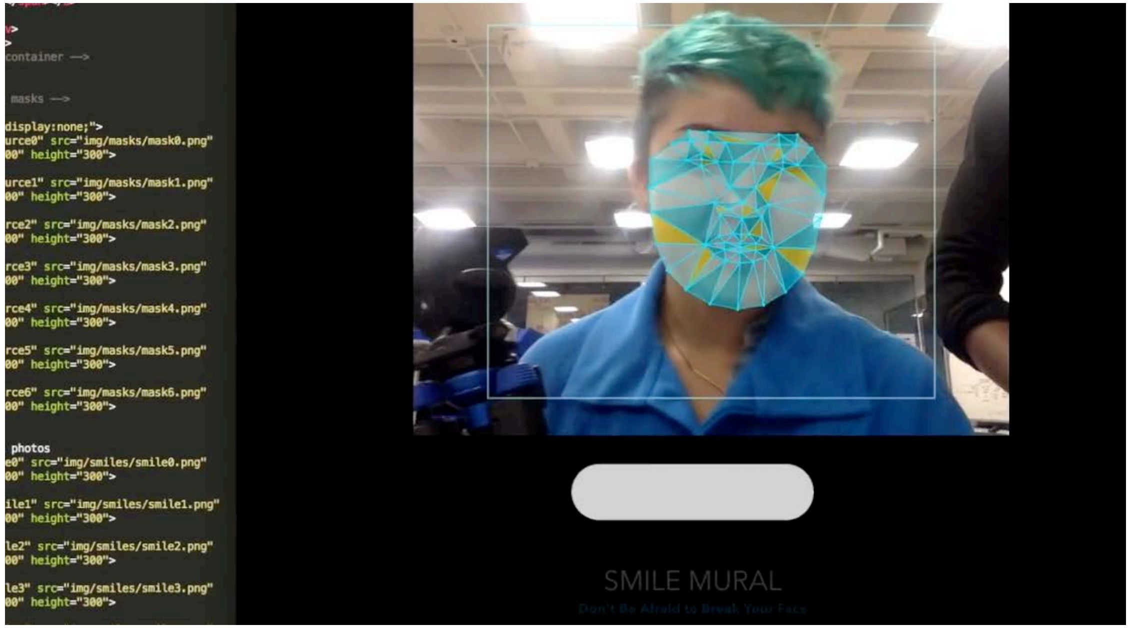
PhD, MIT Pending

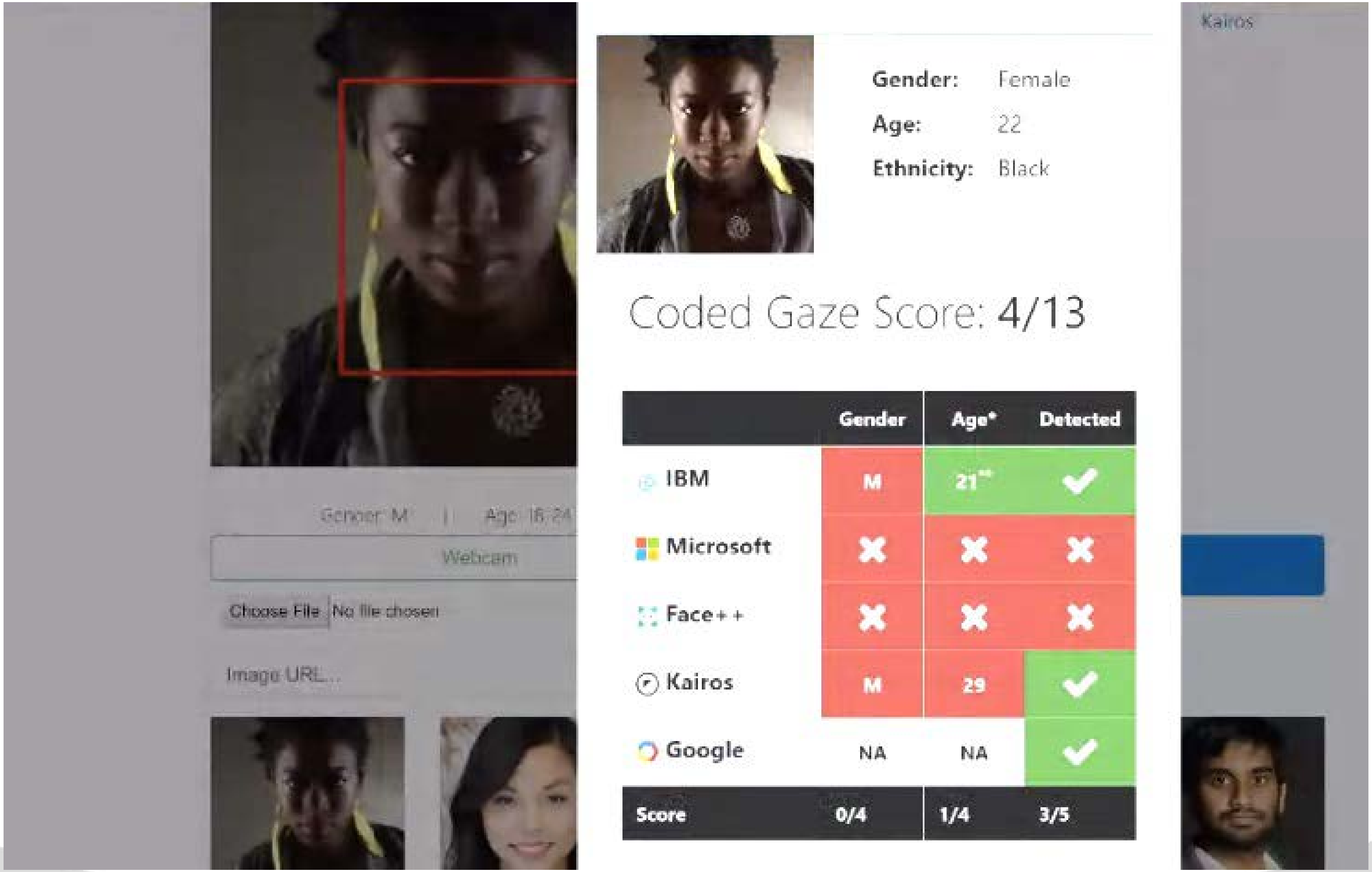
Automated Facial Analysis Tasks



The Coded Gaze

Algorithmic bias creating exclusionary experiences discriminatory practices





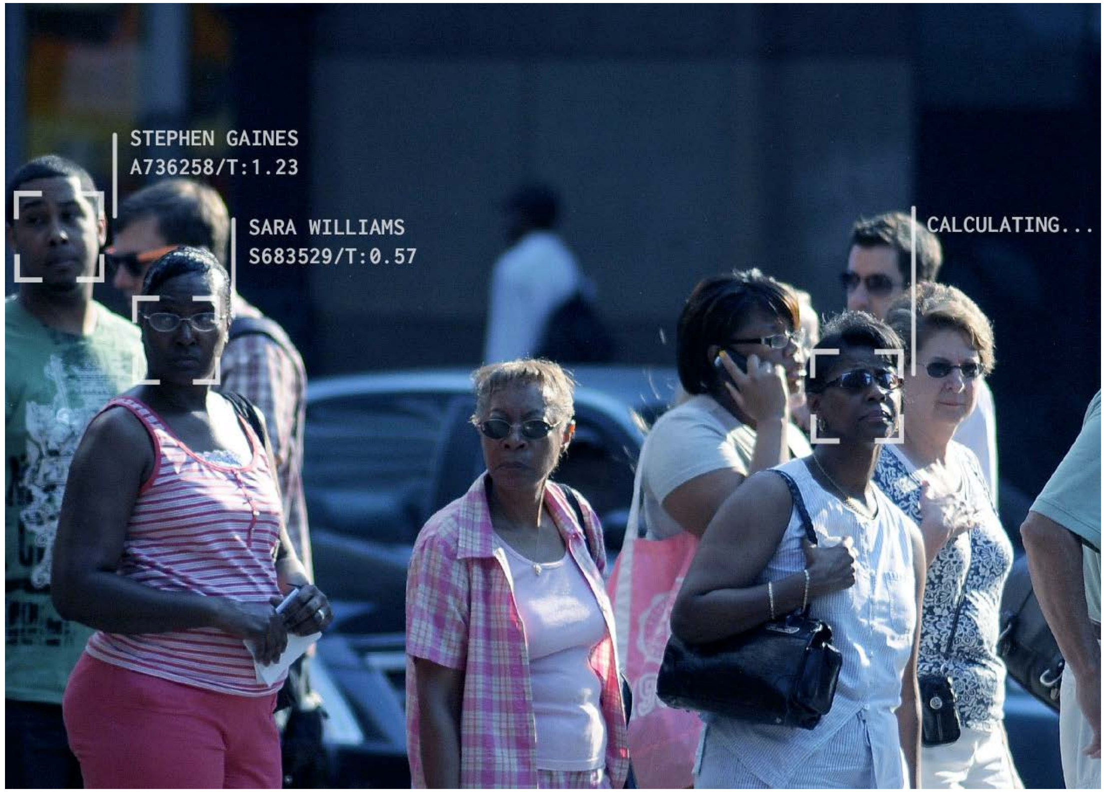
	Gender	Age*	Detected
IBM	M	21**	✔
Microsoft	✖	✖	✖
Face++	✖	✖	✖
Kairos	M	29	✔
Google	NA	NA	✔
Score	0/4	1/4	3/5

Silent Sweep: Over 117 Million US Adults in Face Surveillance Databases

One in two American adults is in a law enforcement face recognition network used in unregulated searches employing algorithms with unaudited accuracy.

The Perpetual Line Up

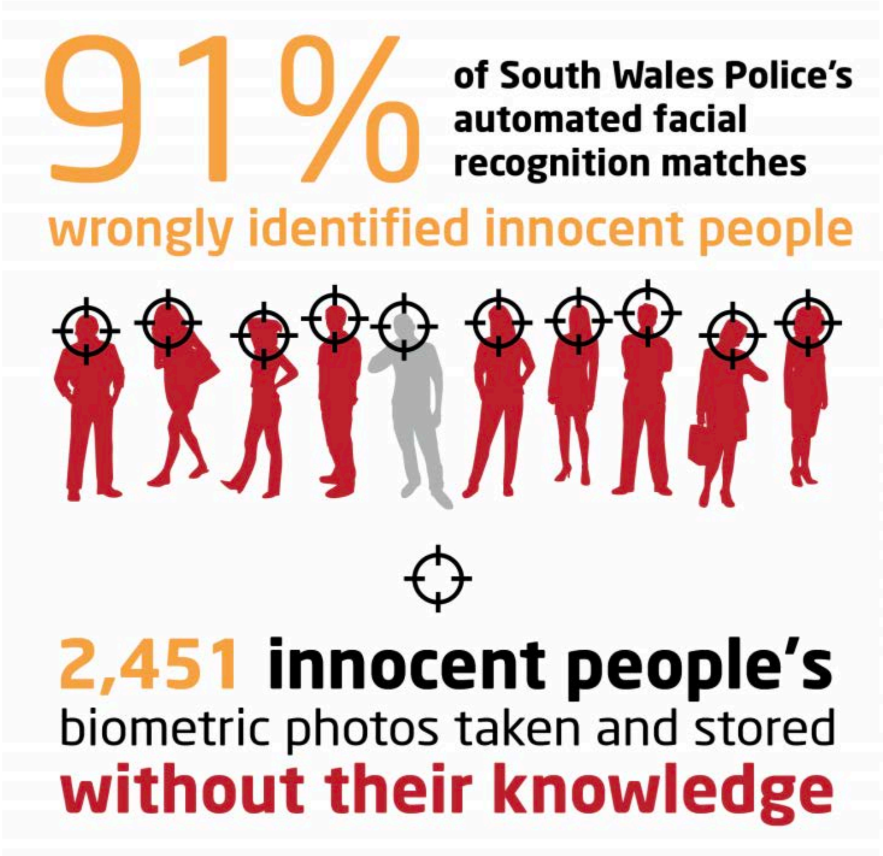
(Garvie , Bedoya, Frankle 2016)



Real-World Impact

"In two cases [Scotland Yard Report], innocent women were matched with men."

- Ian Drury, The Daily Mail – May 15 2018



Expanding Use of Technology

The Hidden Dangers Of Facial Analysis

Joy Buolamwini

WHEN I was a college student using A.I.-powered facial detection software for a coding project, the robot I programmed couldn't detect my dark-skinned face. I had to borrow my white roommate's face to finish the assignment. Later, working on another project as a graduate student at the M.I.T. Media Lab, I resorted to wearing a white mask to have my presence recognized.

My experience is a reminder that artificial intelligence, often heralded for its po-

my area of research and one focus of my work with the Algorithmic Justice League, a group that highlights bias in coding — Google's photo application labeling black people in images as "gorillas" and facial analysis software that works well for white men but less so for everyone else are infamous examples. As disturbing as they are, they do not fully capture the risks of this technology that is increasingly being used in law enforcement, border control, schools' security systems and hiring.

The products of a company called HireVue, which are used by over 600 companies including Nike, Unilever and even Atlanta's public school system, allow em-

repeatedly test the technology and its application.

The tests that have been done on facial analysis technology raise concerns. In collaboration with the computer vision expert Timnit Gebru, I investigated the accuracy of facial analysis technology from IBM, Microsoft and Face++. On the simple task of guessing the gender of a face,

A.I. technology meant to take bias out of hiring could backfire.

all companies' technology performed better with male faces than on female faces and especially struggled with the faces of dark-skinned African women. In the worst case, the technology was 34 percent less accurate for those women than it was for white men.

Companies using HireVue, if they hope to increase fairness, should check their systems to make sure they are not amplifying the biases that informed previous in hiring decisions.

The risks of biased facial analysis technology extend beyond hiring. According to the Center on Privacy and Technology at Georgetown Law, the faces of half of all adults in the United States — over 117 million people — are in face recognition database networks that can be searched by police departments without a warrant. These searches are often reliant on facial recognition technology that hasn't been tested for accuracy on different groups of people, which risks subjecting innocent people to police scrutiny or erroneous criminal charges.

We need to challenge the growing use of this technology, and there has been some progress on this front. The A.C.L.U. is calling on Amazon to stop selling facial analysis technology to law enforcement and is contesting the use of in-car facial recognition for the Vehicle Face System being tested at the United States-Mexico border. Though lawmakers in Texas, Illinois and California have made legislative efforts to regulate facial recognition technology, there are no federal laws. Yet, there is a blueprint. A 2016 report from Georgetown Law School proposed model federal legislation. Policymakers should embrace it.

We can also learn from international models. Unlike the United States, Canada has a federal statute governing the use of

Potential Harms Index

INDIVIDUAL HARMS		COLLECTIVE SOCIAL HARMS
ILLEGAL DISCRIMINATION	UNFAIR PRACTICES	
HIRING		LOSS OF OPPORTUNITY
EMPLOYMENT		
INSURANCE & SOCIAL BENEFITS		
HOUSING		
EDUCATION		
CREDIT		ECONOMIC LOSS
DIFFERENTIAL PRICES OF GOODS		
LOSS OF LIBERTY		SOCIAL STIGMATIZATION
INCREASED SURVEILLANCE		
STEREOTYPE REINFORCEMENT		
DIGNATORY HARMS		

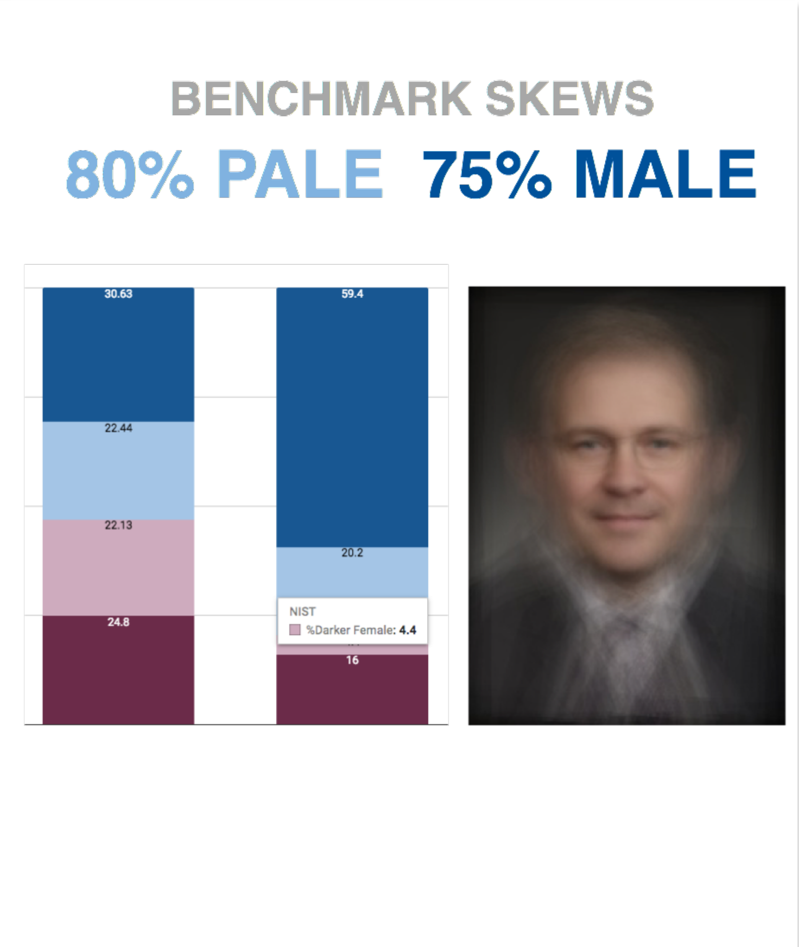
Gender Shades

Intersectional Accuracy Disparities in Commercial Gender Classification

230+ articles in 37+ countries on MIT Thesis Research findings

Buolamwini, J., Gebru, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Gold Standard Measures of Success Mislead



False Sense of Progress

2014
DEEPFACE

97.35%
ACCURACY ON GOLD STANDARD LFW BENCHMARK

(Taigman et al., 2014)

GOLD STANDARD SKEWS
Labeled Faces in The Wild
Released in 2007

~77.5% Male
~83.5% White

(Han and Jain, 2014)

National Benchmarks Not Immune

NIST 2015 IJB-A BENCHMARK

INTERSECTIONAL BREAKDOWN

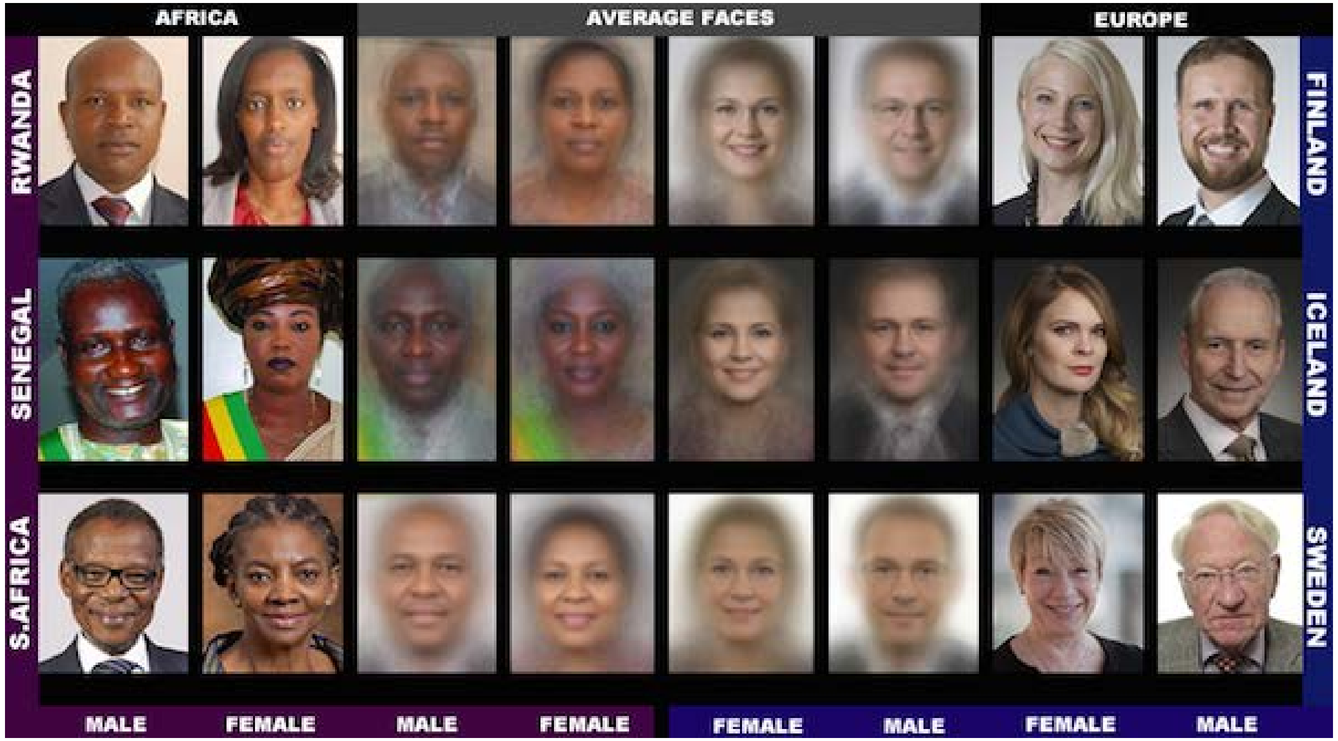
4.4% Darker Female
20.2% Lighter Female

59.4% Lighter Male
16% Darker Male

SINGLE AXIS

24.6% Female
75.4% Male

Towards Better Evaluation



PILOR PARLIAMENTS BENCHMARK

FIRST GENDER AND SKIN TYPE LABELED GENDER CLASSIFICATION BENCHMARK

***54.4%** Male*

***53.6%** Lighter*

Testing Commercial AI Systems

How accurate are systems from IBM, Microsoft, and Face++ at determining the gender of faces in inclusive benchmark?

IBM.WATSON

Face++ 旷视

Microsoft
Cognitive Services

Overall Accuracy

Aggregate performance metrics can mask racial and gender bias

FACE++

IBM

93.7% **90%** **87.9%**

www.gendershades.org

May 2017 PPB Results

Gender Bias

All companies perform better on men than women

	FEMALE FACES	MALE FACES
	89.3%	97.4%
FACE++	78.7%	99.3%
IBM	79.7%	94.4%

FEMALE

8-21% ERROR GAP

MALE

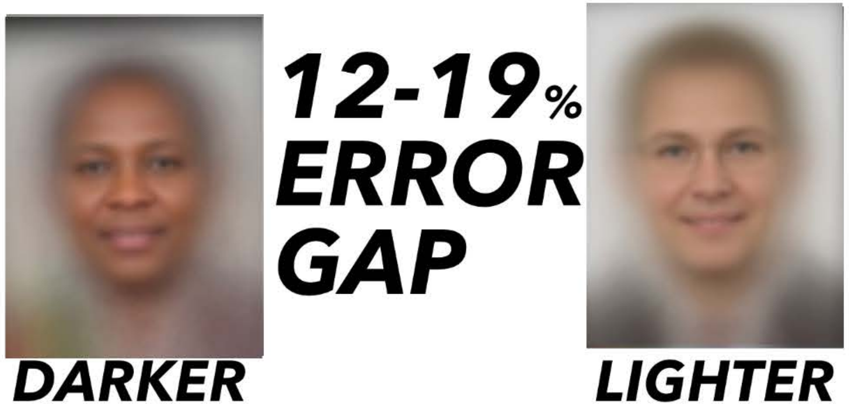
www.gendershades.org

May 2017 PPB Results

Skin Type ~ Racial Bias

All companies perform better on whites than people of color

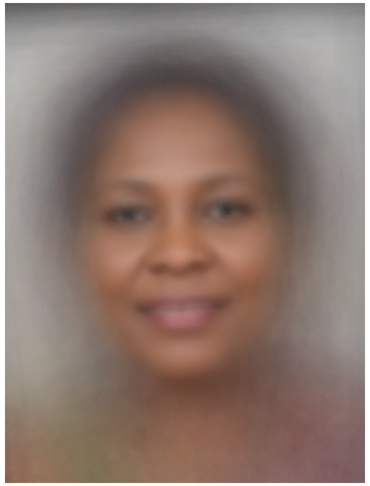
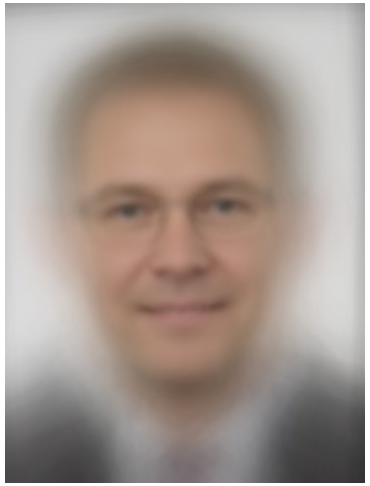
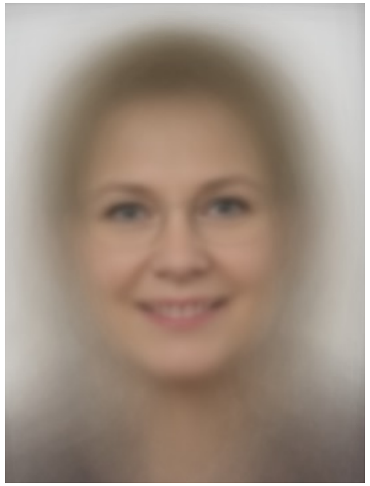
	DARKER FACES	LIGHTER FACES
	87.1%	99.3%
FACE++	83.5%	95.3%
IBM	77.6%	96.8%



www.gendershades.org

May 2017 PPB Results

Intersectional Performance

94%	79.2%	100%	98.3%
			
DARKER MALES	DARKER FEMALES	LIGHTER MALES	LIGHTER FEMALES

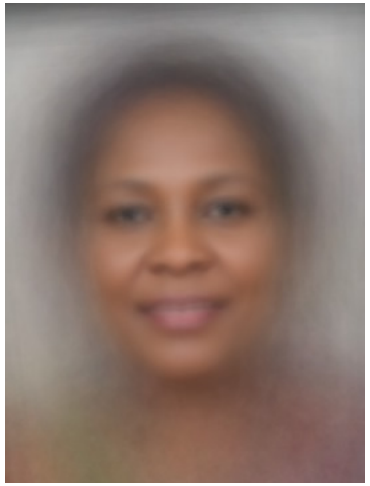
May 2017 PPB Results

Intersectional Performance

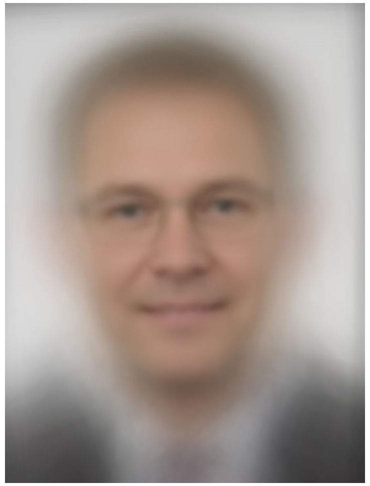
FACE++

99.3%	65.5%	99.2%	94.0%

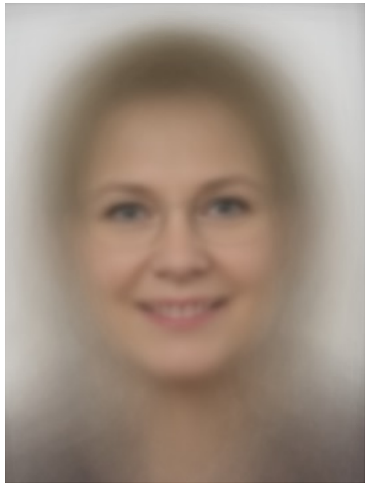
DARKER MALES



DARKER FEMALES



LIGHTER MALES



LIGHTER FEMALES

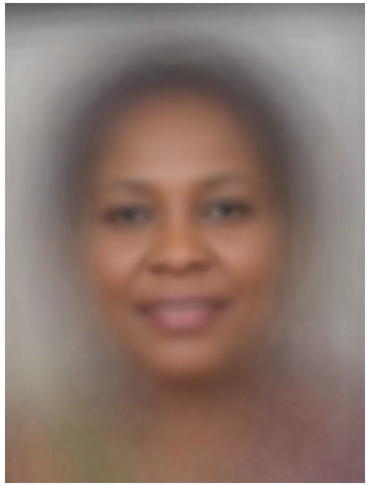
May 2017 PPB Results

Intersectional Performance

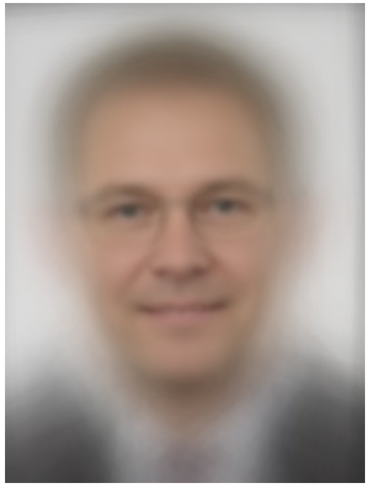
88% | 65.3% | 99.7% | 92.9%

IBM

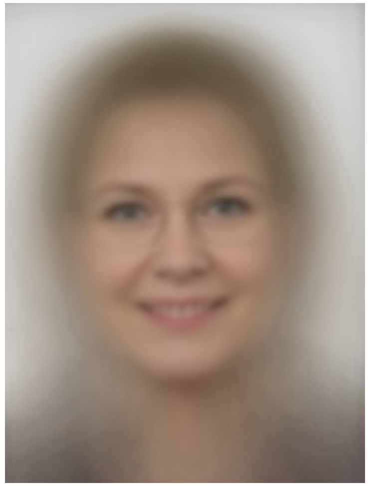
DARKER MALES



DARKER FEMALES



LIGHTER MALES



LIGHTER FEMALES

May 2017 PPB Results

Further Disaggregation Uncovers Even Higher Error Rates

	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
	1.7%	**1.1%**	**3.3%**	**0%**	**23.2%**	**25.0%**
FACE++	**11.9%**	**9.7%**	**8.2%**	**13.9%**	**32.4%**	**46.5%**
IBM	**5.1%**	**7.4%**	**8.2%**	**8.3%**	**33.3%**	**46.8%**

**Commercial Error Rates Per Skin Type on Female Labeled Faces in PPB

May 2017 PPB Results

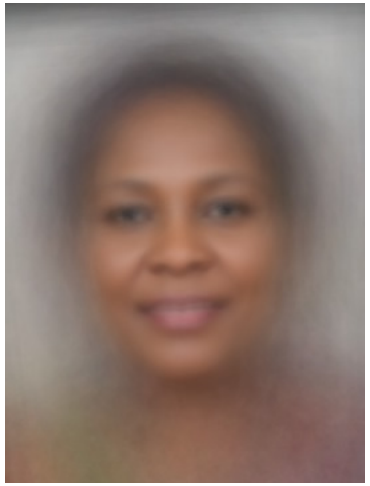
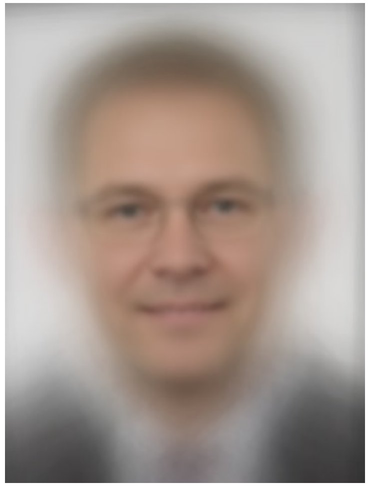
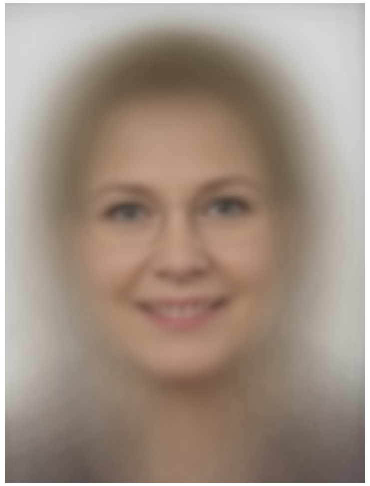
Company Responses to Gender and Racial Bias in Commercial AI Systems

IBM and Microsoft engaged researchers

All companies released new products within 7 months of receiving audit results

Self-Reported Improvement

February 2018 Internal IBM Results

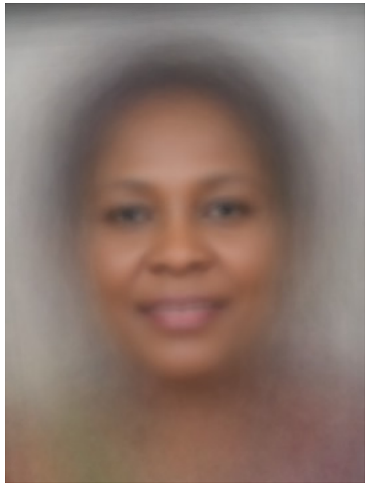
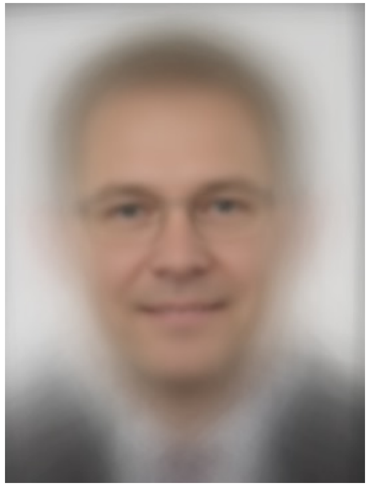
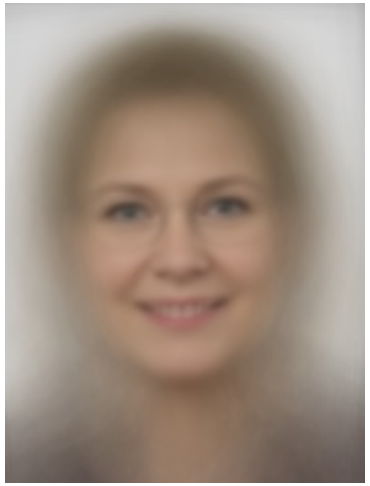
98%	96.5%	99.8%	100%
			
DARKER MALES	**DARKER FEMALES**	**LIGHTER MALES**	**LIGHTER FEMALES**

Self-Reported Results With .99 Treshhold

External Follow-Up Evaluation

August 2018 PPB Results

IBM

99.4%	83.0%	99.7%	97.6%
			
DARKER MALES	DARKER FEMALES	LIGHTER MALES	LIGHTER FEMALES

Accuracy Determined Using Gender Label Returned By API

Accuracy Doesn't Mitigate Abuse

Illustration: Sally Thurer for The Intercept/Getty Images

21

IBM USED NYPD SURVEILLANCE FOOTAGE TO DEVELOP TECHNOLOGY THAT LETS POLICE SEARCH BY SKIN COLOR

Regulators Mitigate Abuse

Gender Shades Research Supported Recommendations

- Require Vendors of Facial Analysis Technology To:
  - Implement internal bias evaluation, mitigation, and reporting procedures
  - Regularly report performance on national benchmarks
  - Support independent evaluation from research community

- Require National Institute of Standards & Technology To:
  - Make public demographic and phenotypic composition of benchmarks
  - Report accessible intersectional performance metrics

Regulators Mitigate Abuse

Broader Considerations

- **Consent and Control:** Ensure consumers have meaningful opportunity to consent or refuse capture of face and ability to control use of face data – (Require companies like Facebook Provide Face Purge Option)

- **Transparency:** Require disclosure when facial analysis technology is in use and information about storage and use of face data

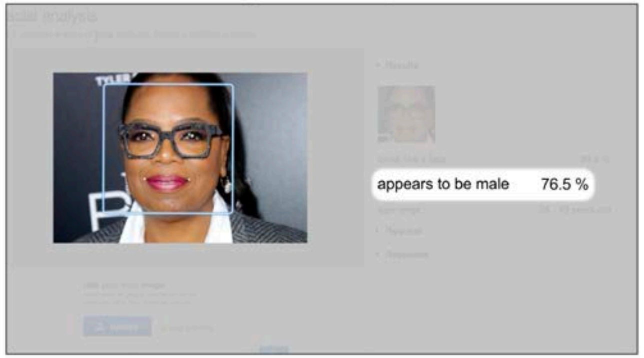
- **Due Process:** Provide mechanisms for redress and contestation of decisions made with or informed by facial analysis technology

- **Heightened Privacy:** Recognize that face images are identifying information, and enable processors to determine consumers' precise geolocation information

For More Information Contact

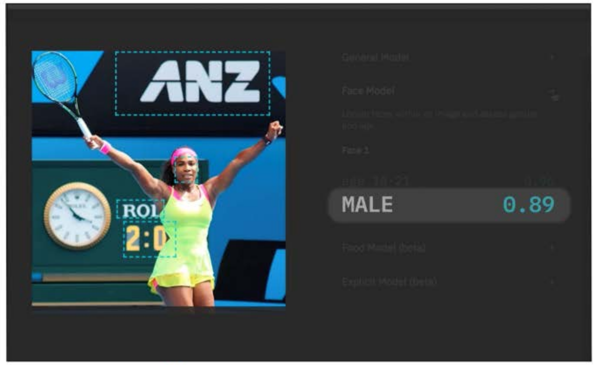
comms@ajlunited.org

Oprah Winfrey



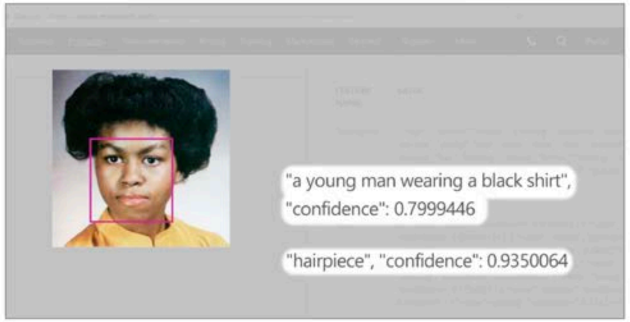
amazon

Serena Williams



IBM WATSON

Michelle Obama



Microsoft

Lunch

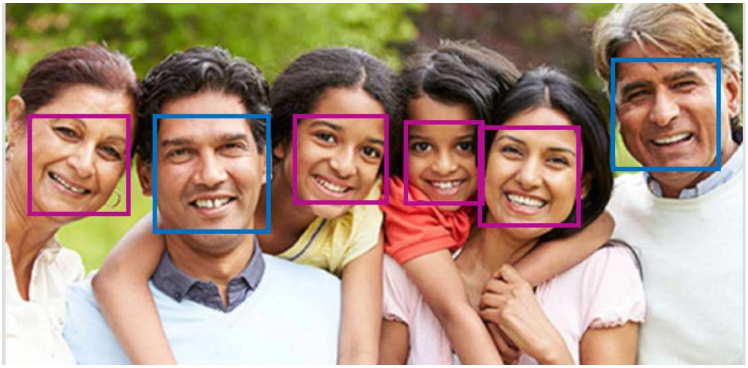
1:00-2:15 pm

Fairness and Intelligibility in Machine Learning Systems

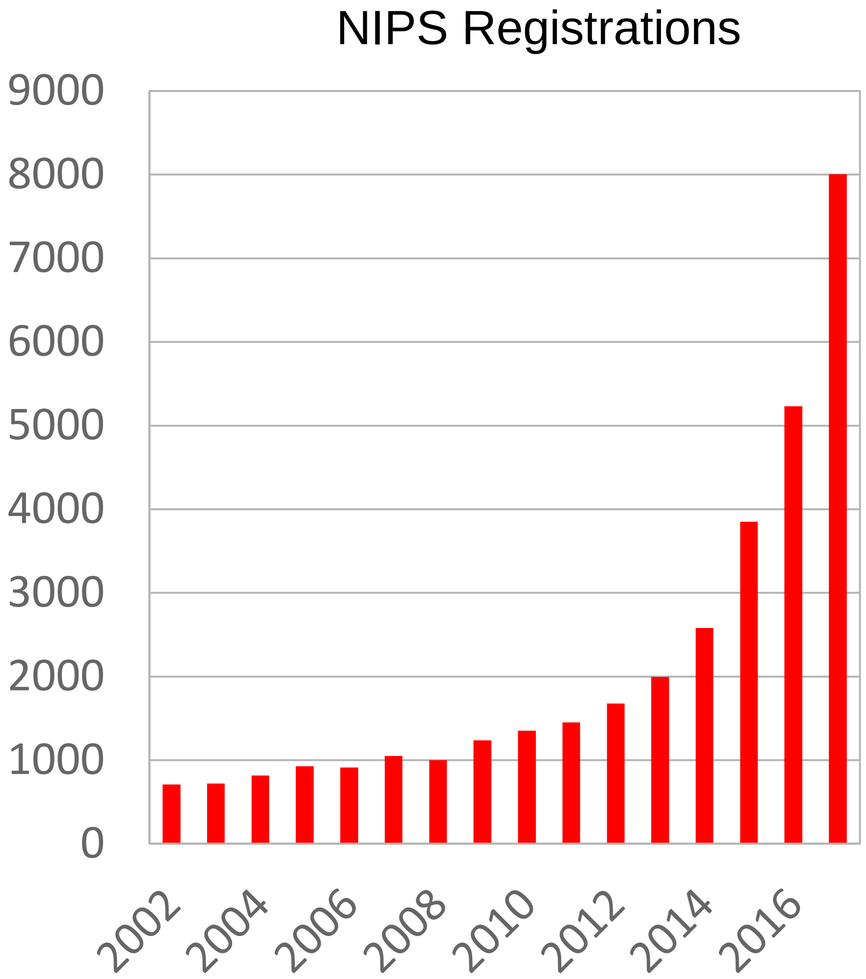
Jenn Wortman Vaughan

Microsoft Research

The Age of AI




skype


NIPS Registrations
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
2002
2004
2006
2008
2010
2012
2014
2016

New Challenges

Online Ads for High-Paying Jobs Are Targeting Men More Than Women New study uncovers gender bias

Do Google's 'unprofessional hair' results show it is racist?
Leigh Alexander

When Algorithms Discriminate

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

When it Comes to Policing, Data Is Not Benign

e searches for 'unprofessional hair for work' (left) and 'professional hair for work' (right) on oogle

Amazon just showed us that 'unbiased' algorithms can be inadvertently racist

Technology

Google apologises for Photos app racist blunder

1 July 2015 | Technology

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden

Amazon Prime and the racist algorithms

Microsoft’s AI Principles

Fairness

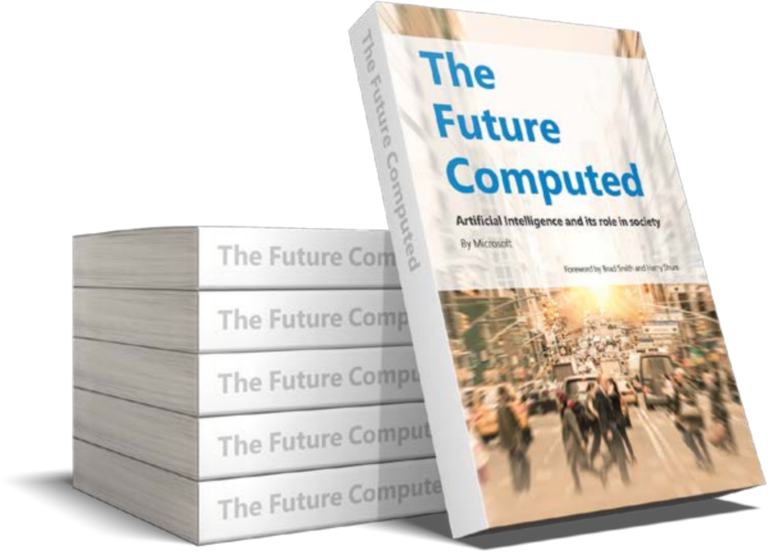
Reliability & Safety

Privacy & Security

Inclusive-ness

Transparency

Accountability



FATE: Fairness, Accountability, Transparency, and Ethics in AI

Sensitive Uses of AI

AI Reliability and Safety

Human-AI Collab and Interaction

Fairness and Bias

Intelligibility & Explainability

Engineering Practices for AI

Human Attention & Cognition

AETHER Committee

AI Ethics and Effects in Engineering and Research

Partnership on AI

to benefit people and society

FOUNDING PARTNERS

What are machine learning and AI?

AI

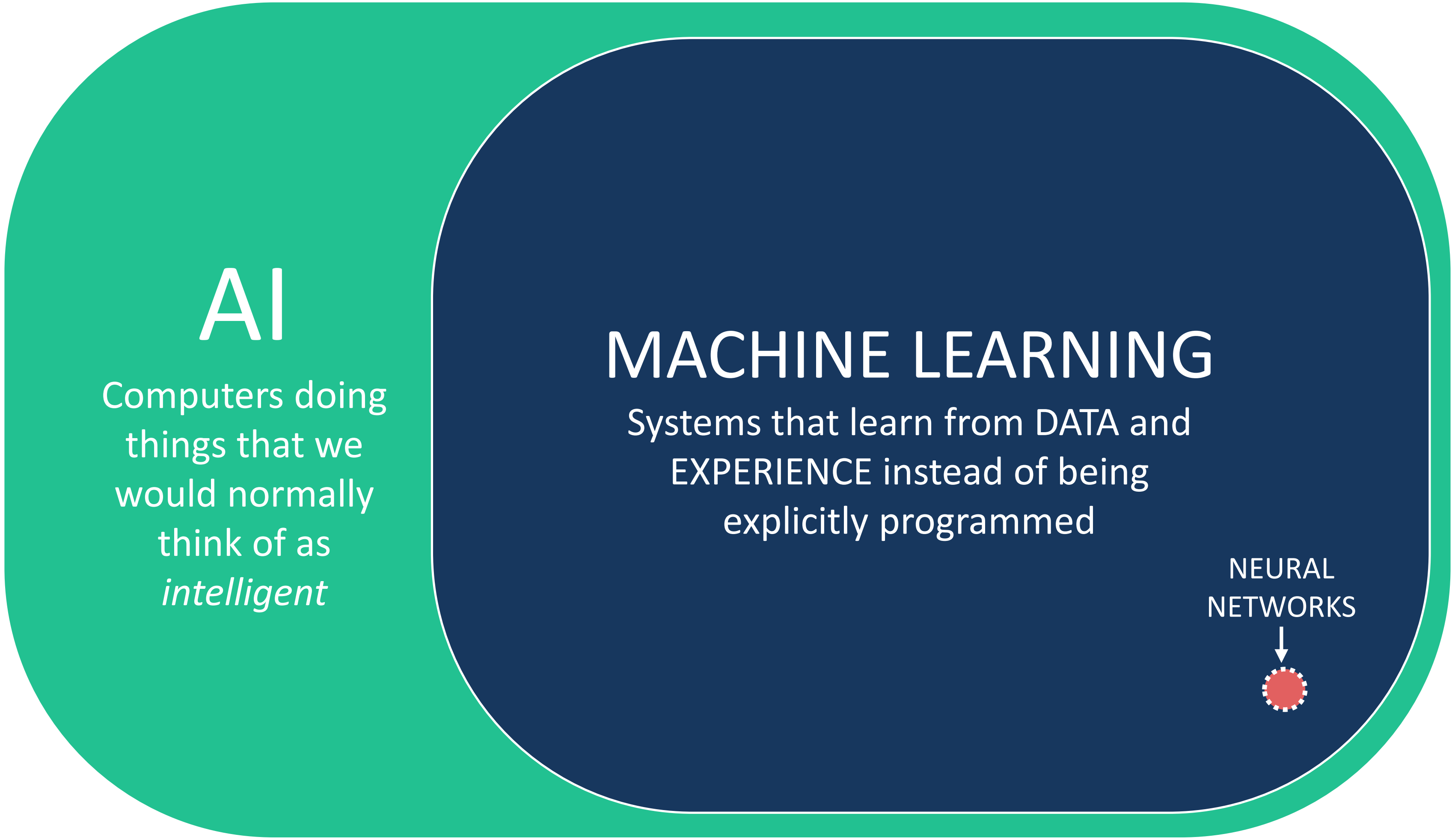
Computers doing things that we would normally think of as *intelligent*

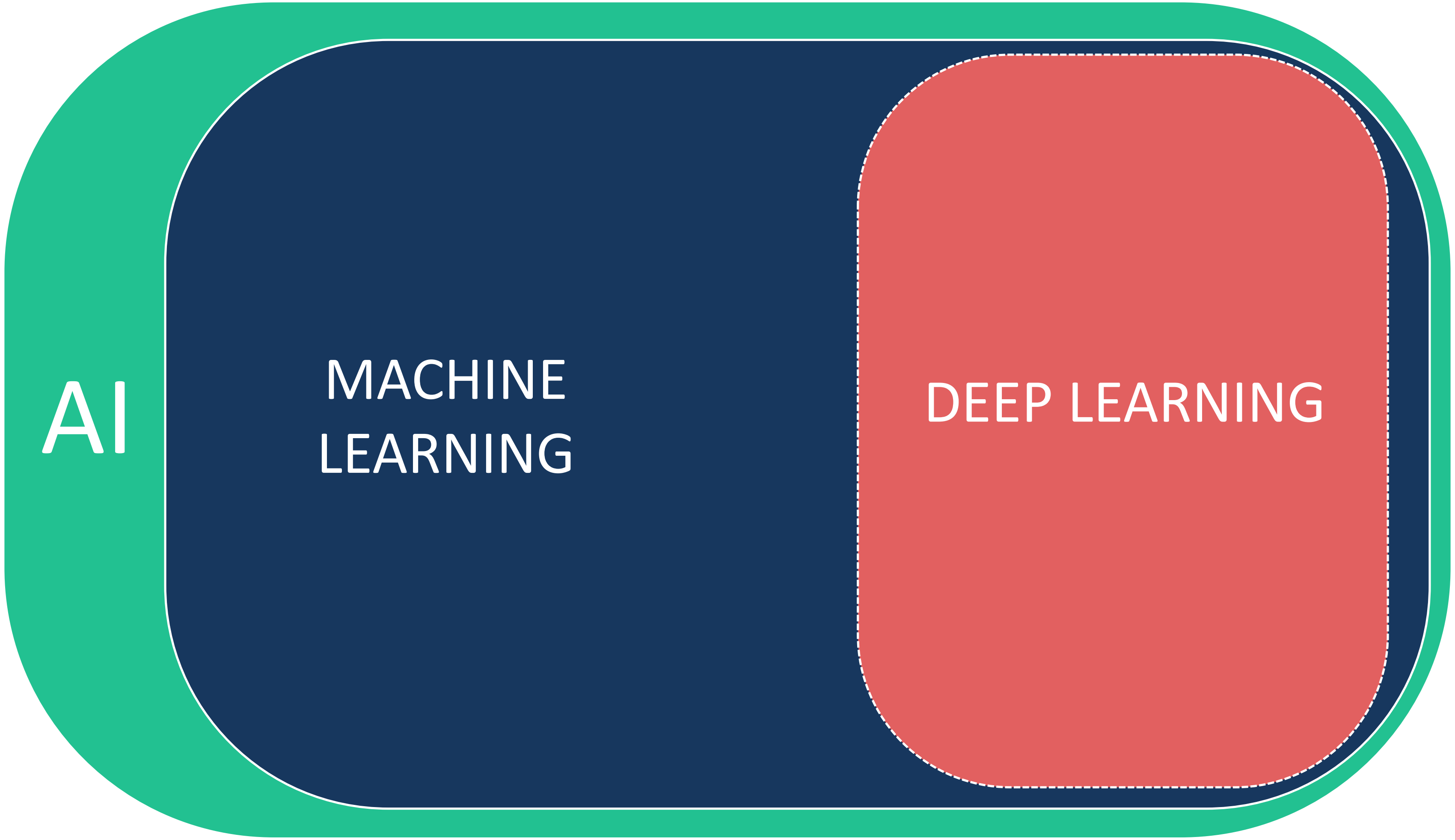
AI

Computers doing things that we would normally think of as *intelligent*

MACHINE LEARNING

Systems that learn from DATA and EXPERIENCE instead of being explicitly programmed


AI
Computers doing things that we would normally think of as *intelligent*
MACHINE LEARNING
Systems that learn from DATA and EXPERIENCE instead of being explicitly programmed
NEURAL NETWORKS

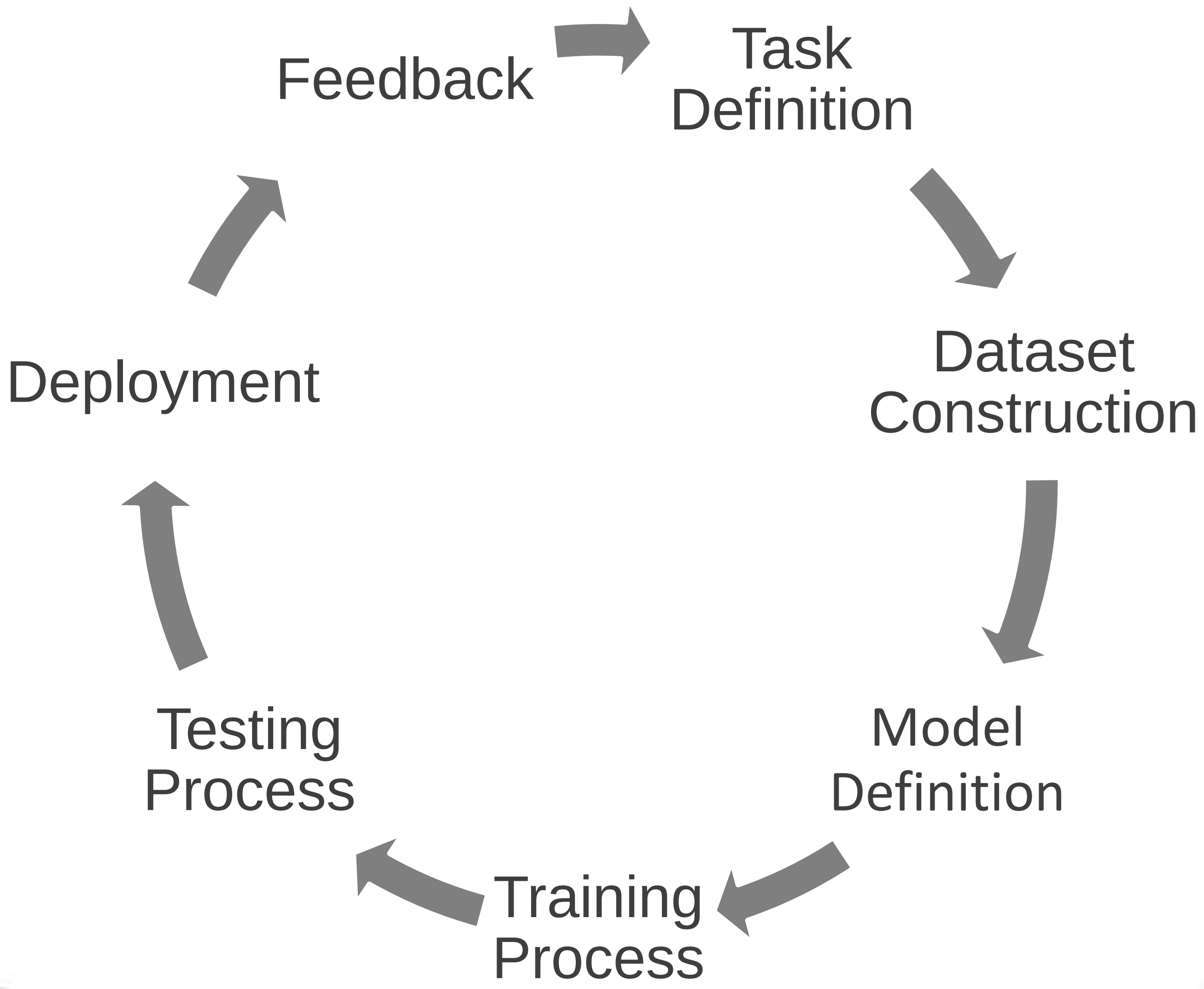

AI
MACHINE LEARNING
DEEP LEARNING

Types of Machine Learning

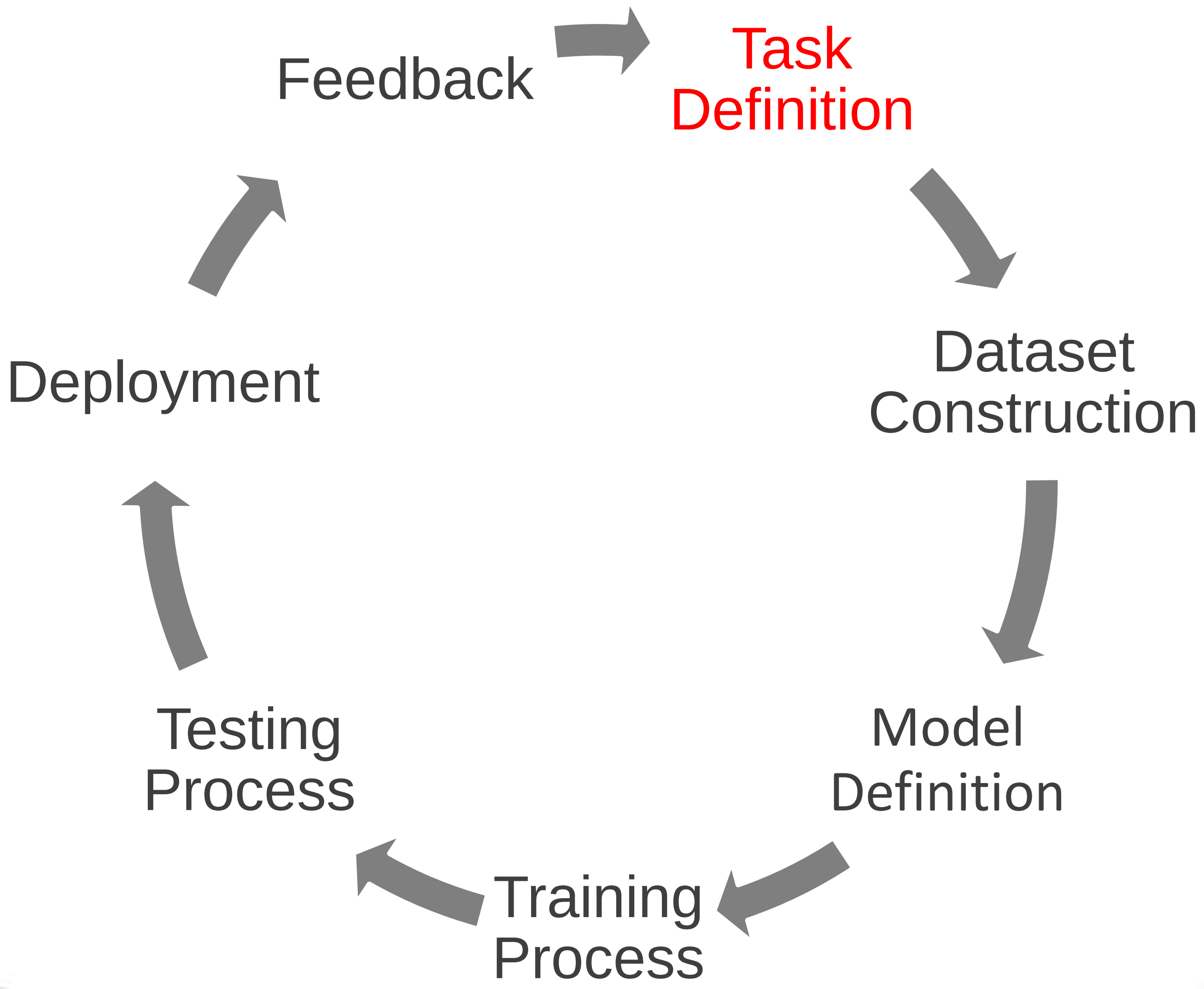
- **Supervised learning:** Use labeled data to learn a general rule mapping inputs to outputs
- **Unsupervised learning:** Identify hidden structure and patterns in data; cluster data points
- **Reinforcement learning:** Perform a task, such as driving a vehicle or playing a game, in a dynamic environment, learning through trial and error

Why might a machine learning system be unfair?

The Machine Learning Pipeline



Task Definition



Task Definition

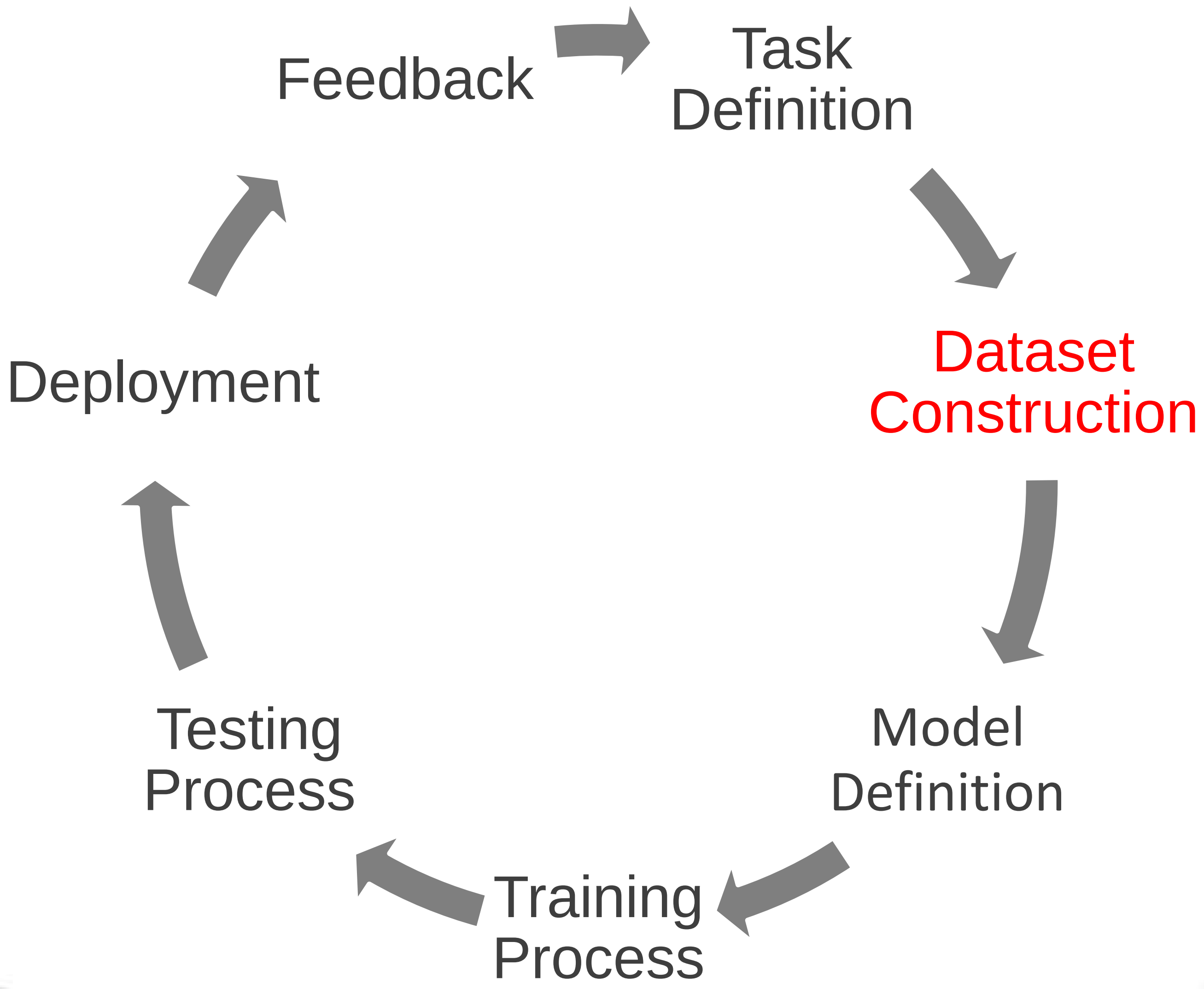
(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

(Wu and Zhang, 2016)

Dataset Construction



Data: Societal Bias

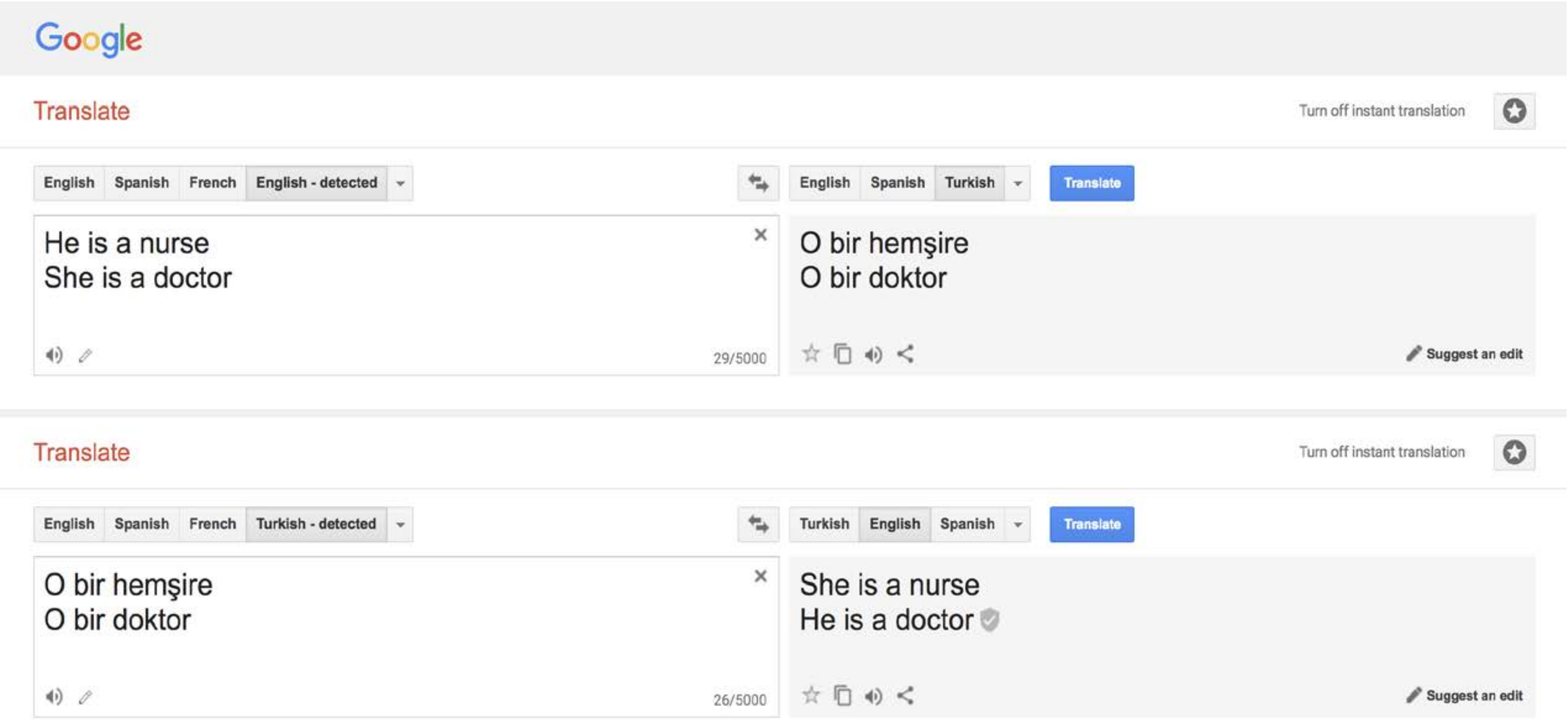
Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

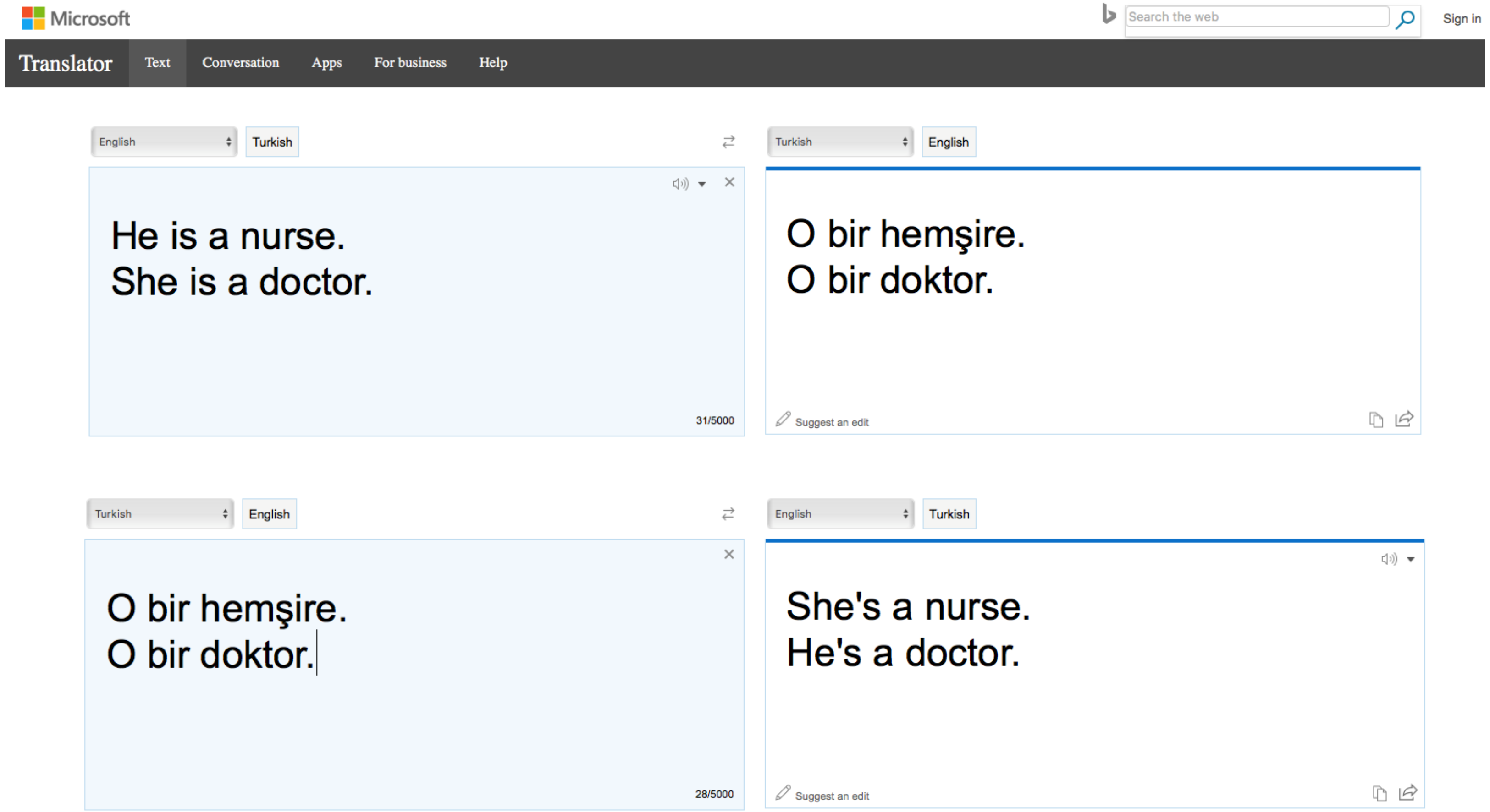
SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Data: Societal Bias



(Caliskan et al., 2017)

Data: Societal Bias



Data: Skewed Sample

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

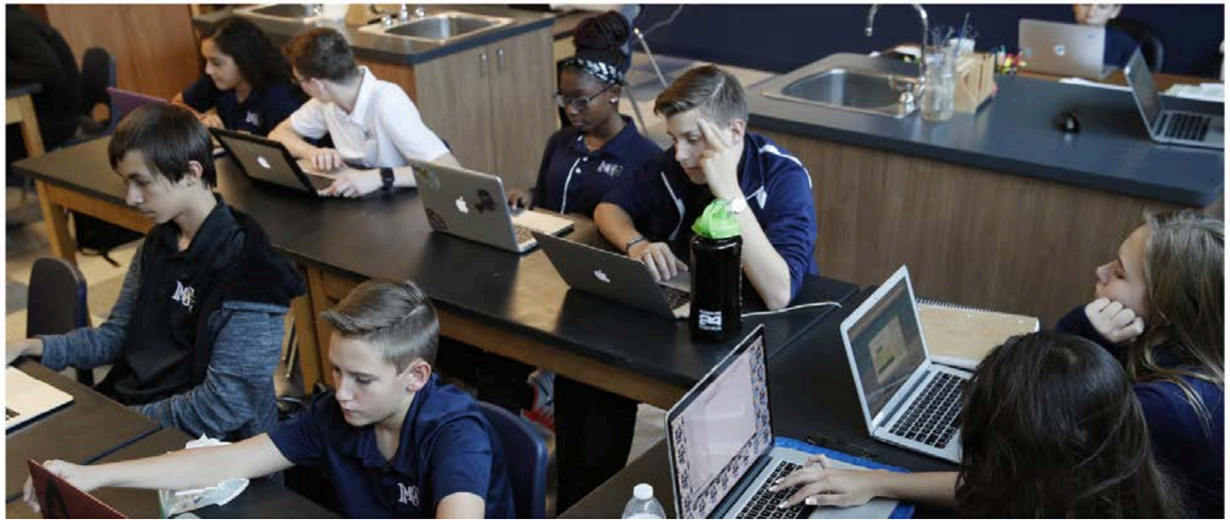
(Buolamwini and Gebru, 2018)

Data: Labeler Bias

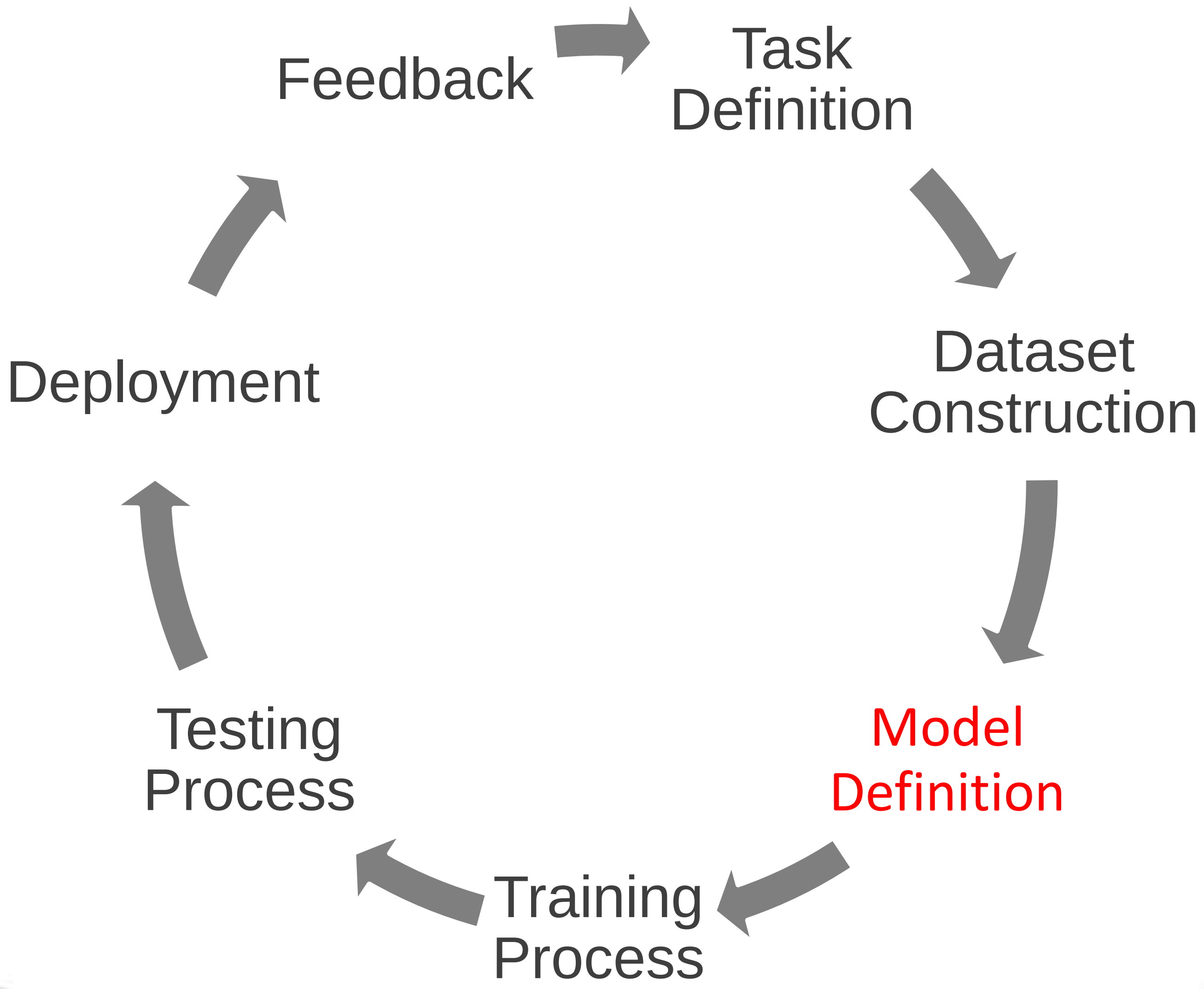
More States Opting To 'Robo-Grade' Student Essays By Computer

June 30, 2018 · 8:13 AM ET
Heard on Weekend Edition Saturday

TOVIA SMITH



Model Definition



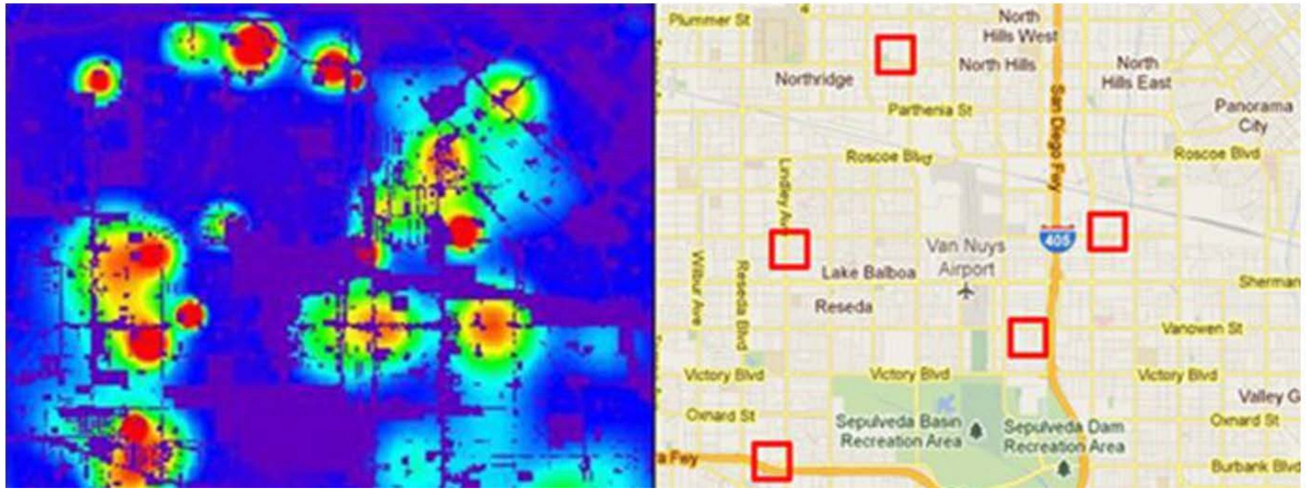
Models are Mathematical Abstractions

$$\begin{aligned}\text{price of house} \quad = \quad & w_1 * \text{number of bedrooms} \\ & + w_2 * \text{number of bathrooms} \\ & + w_3 * \text{square feet} \\ & + \text{a little bit of noise}\end{aligned}$$

Model: Assumptions

Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.



Training Process

Feedback

Task Definition

Dataset Construction

Model Definition

Training Process

Testing Process

Deployment

Training Process

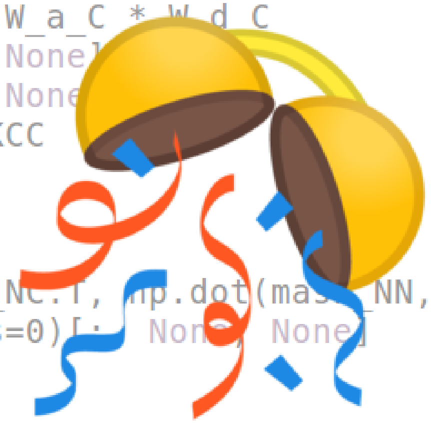
$$
\begin{aligned}
\text{price of house} \quad = \quad & {\color{red}w_1} * \text{number of bedrooms} \\
& + {\color{red}w_2} * \text{number of bathrooms} \\
& + {\color{red}w_3} * \text{square feet} \\
& + \text{a little bit of noise}
\end{aligned}
$$

Training Process

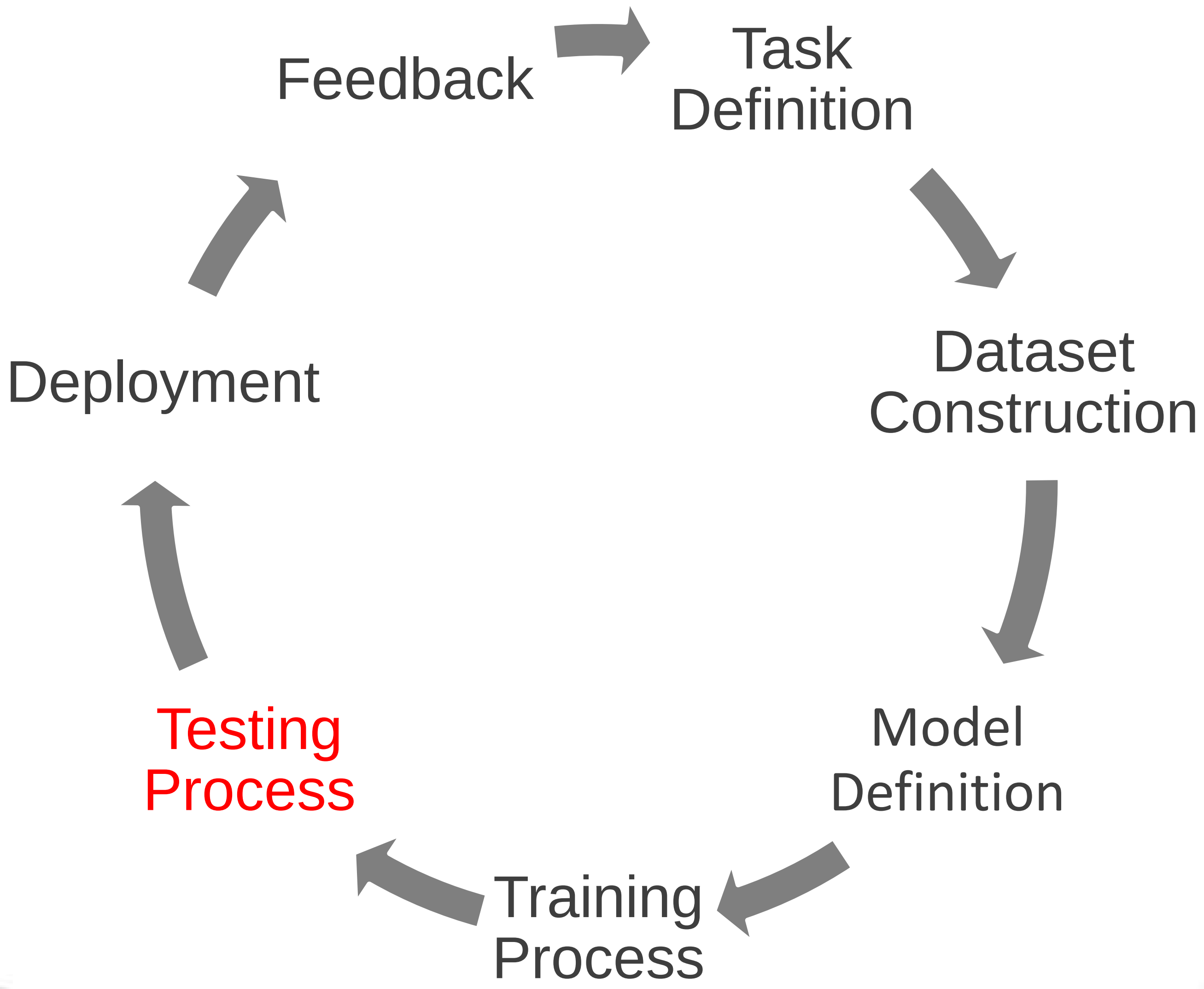
```
if schedule['Lambda_SKCC'] <= self.total_iter:
    start = time.time()

    shp_SKCC[:] = np.outer(W_d_C, W_d_C)
    shp_SKCC[:, :, bool_diag_CC] = W_a_C * W_d_C
    shp_SKCC *= W_K[None, :, None, None]
    shp_SKCC *= W_S[:, None, None, None
    post_shp_SKCC = shp_SKCC + Y_SKCC

    if mask.ndi      2:
        mask_N
        zeta_S       p.    Theta_NC.T, np.dot(mas  NN, Theta_NC))
        zeta_        TS.sum(axis=0)[:, None, None]
    else:
        mask_TNN = mask
        zeta_TNC = np.einsum('tij,jd->tid', mask_TNN, Theta_NC)
        zeta_TCC = np.einsum('tid,ic->tcd', zeta_TNC, Theta_NC)
        zeta_SCC = np.einsum('tcd,ts->scd', zeta_TCC, Psi_TS)
    post_rte_SKCC = d + zeta_SCC[:, None, :, :]
```



Testing Process



Testing: Metrics

Translation tutorial:
21 fairness definitions and their politics

Arvind Narayanan
(Computer scientist, Princeton University)

Computer scientists and statisticians have devised numerous mathematical criteria to define what it means for a classifier or a model to be fair. The proliferation of these definitions represents an attempt to make technical sense of the complex, shifting social understanding of fairness. Thus, these definitions are laden with values and politics, and seemingly technical discussions about mathematical definitions in fact implicate weighty normative questions. A core component of these technical discussions has been the discovery of trade-offs between different (mathematical) notions of fairness; these trade-offs deserve attention beyond the technical community.

Testing: Metrics

	Unqualified	Qualified
Reject	TN	FN
Hire	FP	TP

Testing: Metrics

	Unqualified	Qualified
Reject	TN	FN
Hire	FP	TP

What is the probability that a woman is qualified given that you choose to hire her? What about a man?

Predictive parity requires (almost) equal values of

$$\frac{TP}{TP + FP}$$

Testing: Metrics

	Unqualified	Qualified
Reject	TN	FN
Hire	$\boxed{\text{FP}}$	TP

What is the probability of hiring a woman if she is unqualified? What about a man?

False positive rate balance requires (almost) equal values of

$$\frac{\text{FP}}{\text{FP} + \text{TN}}$$

Testing: Metrics

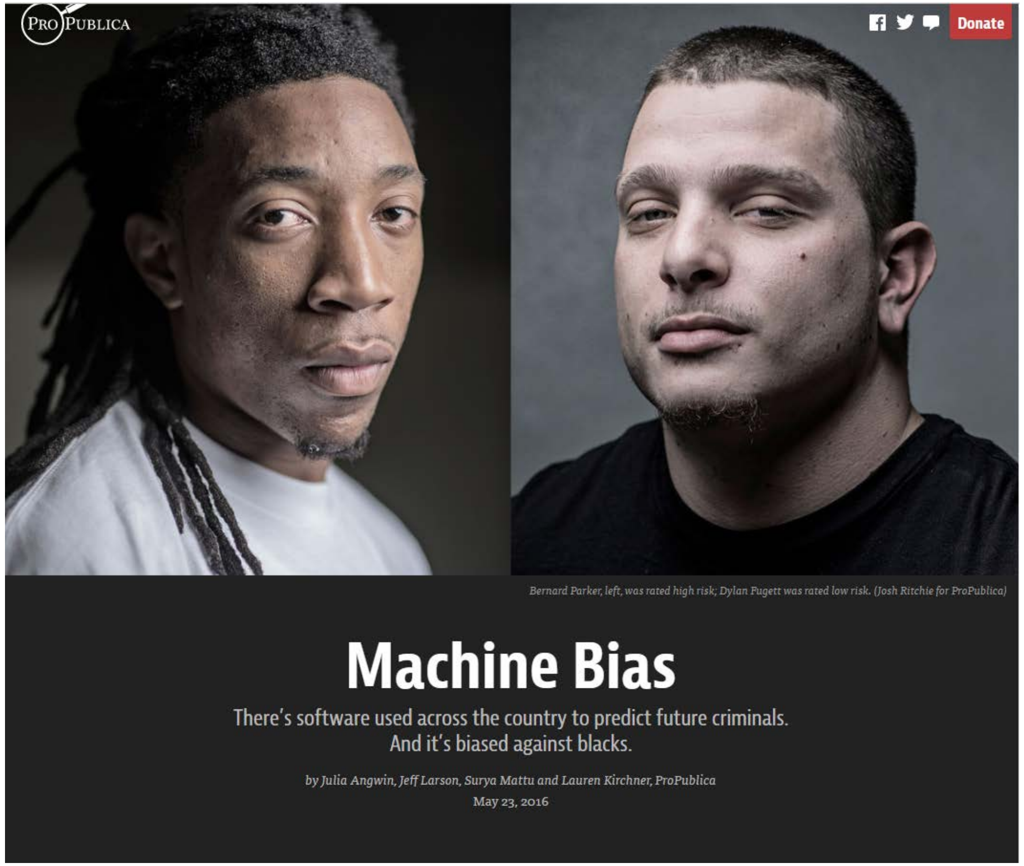
	Unqualified	Qualified
Reject	TN	FN
Hire	FP	TP

What is the probability of rejecting a woman if she is qualified? What about a man?

False negative rate balance requires (almost) equal values of

$$\frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$$

Testing: Metrics



Testing: Metrics

RESPONSE TO PROPUBLICA: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY

The website ProPublica recently published a story that focused on the scientific validity of COMPAS, raising questions about racial bias. As a result of the article and the subsequent national attention that it garnered, Northpointe launched an in-depth analysis of the data samples used by ProPublica. Drawing from the results of our analysis of ProPublica's data, Northpointe unequivocally rejects the ProPublica conclusion of racial bias in the COMPAS risk scales.

Predictive modeling is a specialized field within statistics and the appropriate use and interpretation of valid predictive models require a solid understanding of the techniques and methodological nuances common to this type of work. Our detailed review of how ProPublica conducted their analysis revealed several statistical and technical errors such as misspecified regression models, mis-defined classification terms and measures of discrimination, the incorrect interpretation and use of model errors, and more. These errors led to a false conclusion of racial bias; we do not

Testing: Metrics

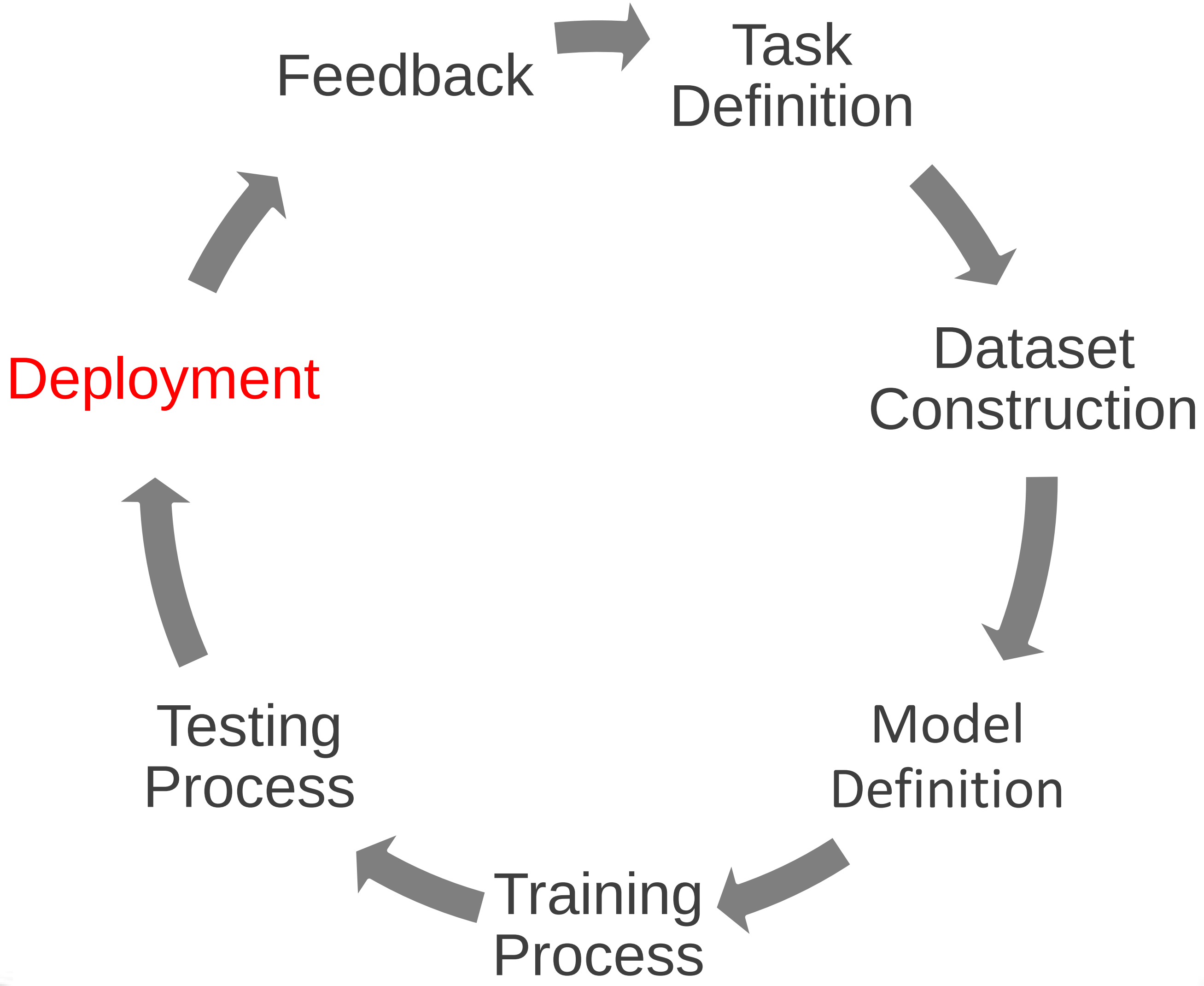
Monkey Cage

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

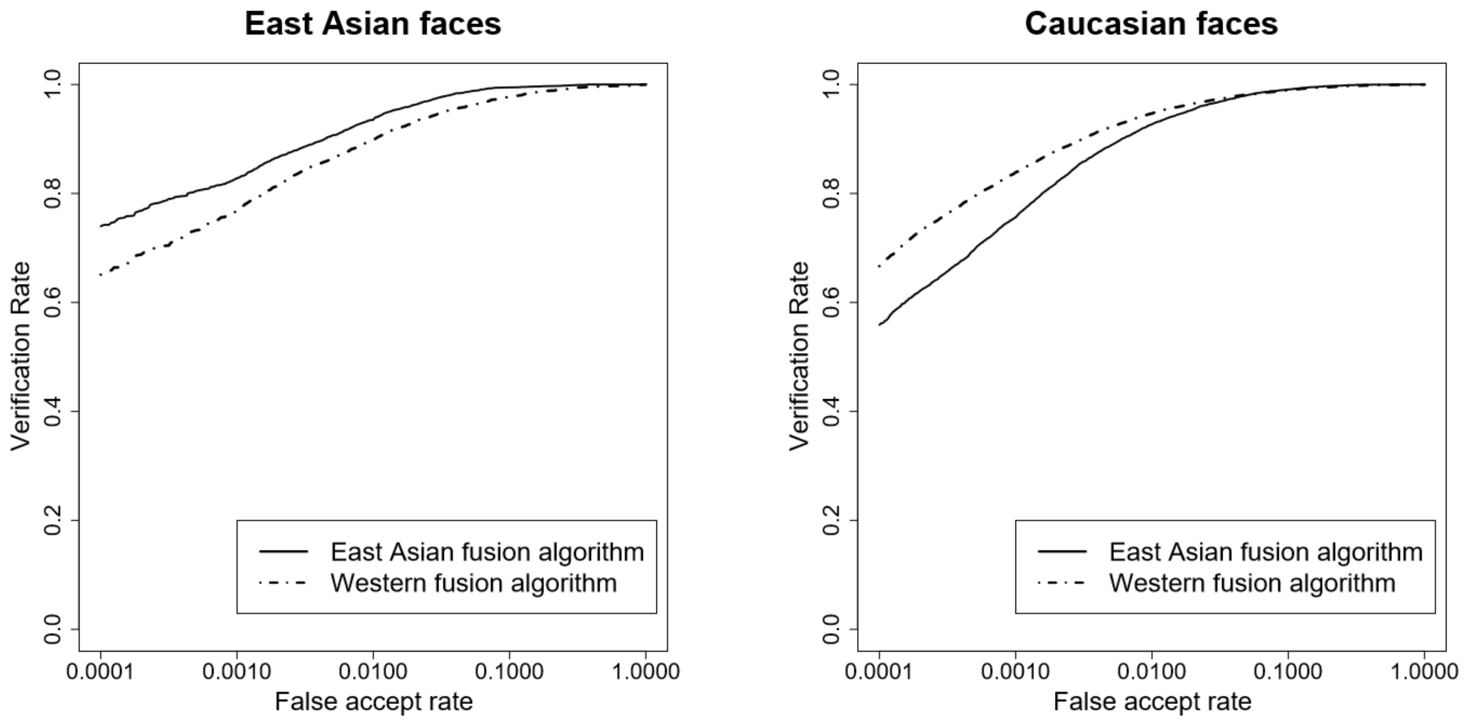
By **Sam Corbett-Davies**, **Emma Pierson**, **Avi Feller** and **Sharad Goel**
October 17, 2016

(Kleinberg et al., 2016; Chouldechova, 2017)

Deployment

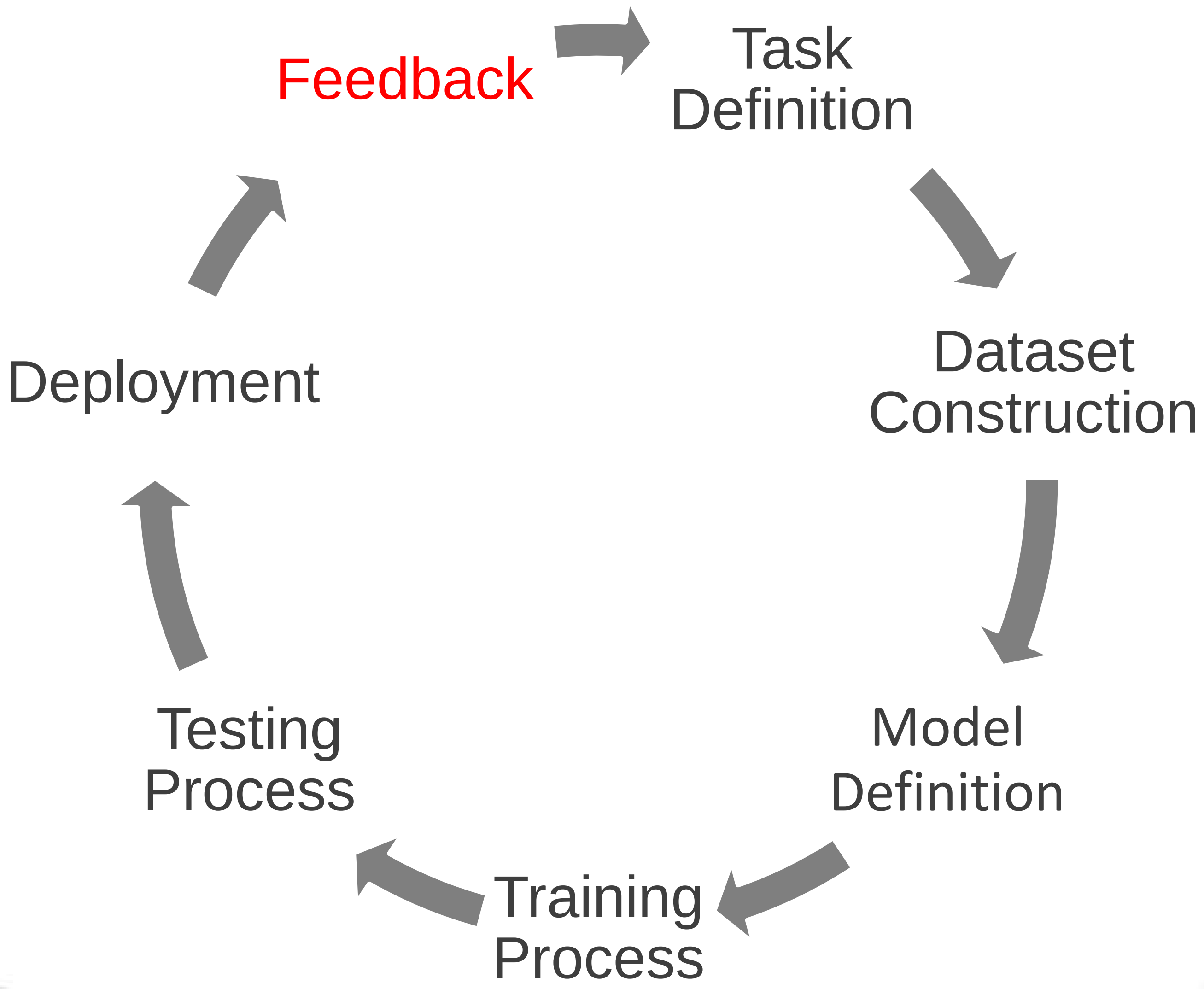

Feedback
Task Definition
Dataset Construction
Model Definition
Training Process
Testing Process
Deployment

Deployment: Context



(Phillips et al., 2011)

Feedback


Feedback
Task Definition
Dataset Construction
Model Definition
Training Process
Testing Process
Deployment

Feedback Loops

Use history of drug-crime reports and arrests to predict future crime locations…

More historic arrests in Black and Hispanic areas

More policing in these areas

More arrests in these areas

So what can we do?

Strategies to Mitigate Harms

- Prioritize fairness at every stage of the ML pipeline
- Think critically about implicit assumptions made at each stage
- Pay attention to potential biases in the data source and data preparation process
- Check if test data matches the deployment context
- Involve diverse stakeholders and gather multiple perspectives
- Acknowledge our mistakes and learn from them

Transparency vs. Intelligibility

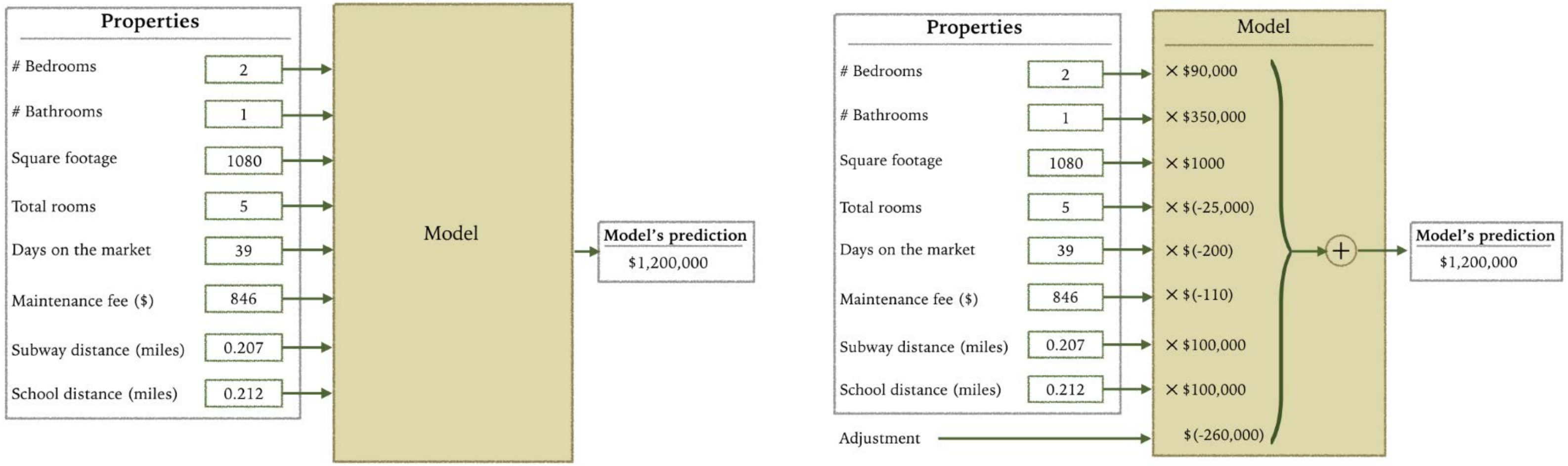
What is Transparency?

- In policy circles, transparency represents two distinct ideas
  - People should be able to understand and monitor how AI systems work
  - Those who deploy AI systems should be honest and forthcoming about how and when they are being used

- In machine learning circles, the former is called “intelligibility” or “interpretability,” and **literal transparency can work against it!**

Transparency ≠ Intelligibility

- Exposing ML source code doesn't tell us much
- Exposing model internals can stop people from noticing when a model makes a mistake because of information overload

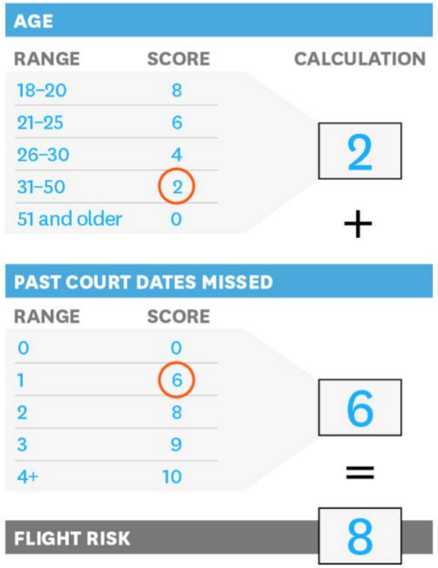


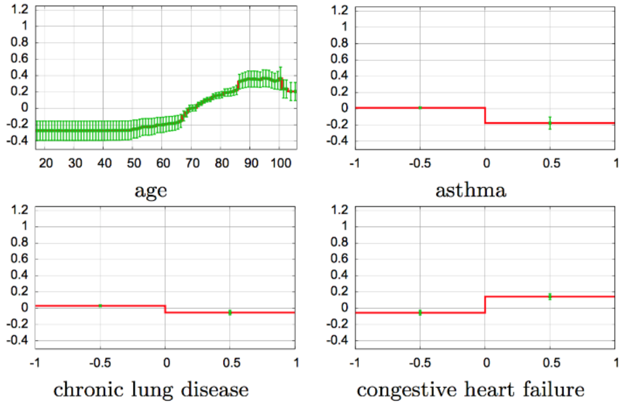
(Poursabzi-Sangdeh et al., 2018)

Why intelligibility?

- Accountability: An applicant wants to know why she was denied a loan.
- Trust: A model deployed in a school predicts that a student is likely to drop out. Knowing the factors relevant for the prediction could help his teacher decide whether to believe it and how to intervene.
- Bias assessment: A model matches candidates to jobs. By understanding characteristics of the training data, an employer may see that female candidates are underrepresented, leading to potential bias.
- Robustness: A data scientist sees unexpected predictions from a model she has trained. Knowing why these predictions were made could help her debug the model.

Intelligibility via “Simple Models”





$$y = f_1(x_1) + \ldots + f_d(x_d)$$

Point Systems

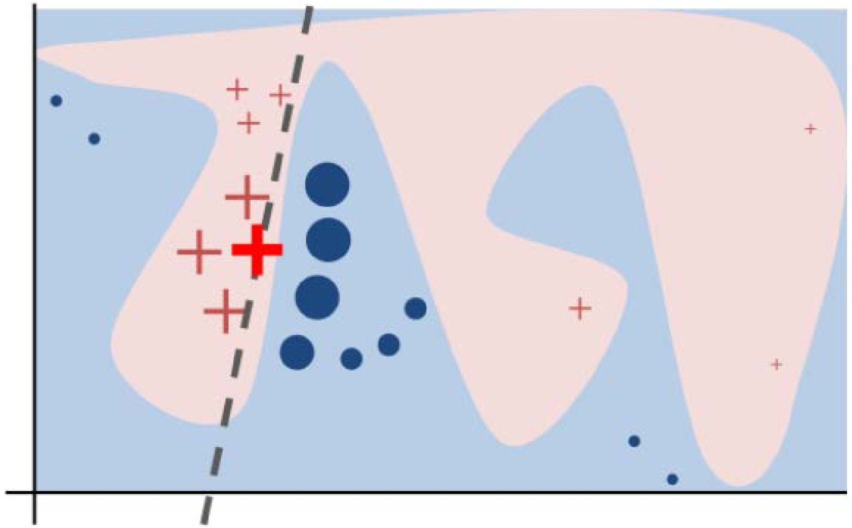
(Jung et al., 2017; Ustun & Rudin, 2015)

Generalized Additive Models

(Lou, Caruana, et al., 2012&2013)

Classic methods: decision trees, rule lists (if-then-else), rule sets, sparse linear models, …

Intelligibility via Post Hoc Explanations



Simple Explanations of a Single Prediction
(e.g., Ribeiro et al., 2016; Lundberg and Lee, 2017)

If Age <50 and Male =Yes:

If Past-Depression =Yes and Insomnia =No and Melancholy =No, then Healthy

If Past-Depression =Yes and Insomnia =Yes and Melancholy =Yes and Tiredness =Yes, then Depression

If Age ≥ 50 and Male =No:

If Family-Depression =Yes and Insomnia =No and Melancholy =Yes and Tiredness =Yes, then Depression

If Family-Depression =No and Insomnia =No and Melancholy =No and Tiredness =No, then Healthy

Default:

If Past-Depression =Yes and Tiredness =No and Exercise =No and Insomnia =Yes, then Depression

If Past-Depression =No and Weight-Gain =Yes and Tiredness =Yes and Melancholy =Yes, then Depression

If Family-Depression =Yes and Insomnia =Yes and Melancholy =Yes and Tiredness =Yes, then Depression

Simple Approximations of a Full Model
(e.g., Lakkaraju et al., 2017)

Data Intelligibility: Datasheets for Datasets

Motivation for Dataset Creation

Why was the dataset created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

What (other) tasks could the dataset be used for?

The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.[2]

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

Who funded the creation of the dataset?

The building of the LFW database was supported by a United States National Science Foundation CAREER Award.

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

How many instances are there? (of each type, if appropriate)?

The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

[1] All information in this datasheet is taken from one of five sources. Any errors that were introduced from these sources are our fault.

Original paper: http://www.cs.cornell.edu/people/pabo/movie-review-data/; LFW survey: http://vis-www.cs.umass.edu/lfw/lfw.pdf; Paper measuring LFW demographic characteristics : http://biometrics.cse.msu.edu/Publications/Face/HanJain.UnconstrainedAgeGenderRaceEstimation.MSUTechReport2014.pdf; LFW website: http://vis-www.cs.umass.edu/lfw/.

[2] Unconstrained face recognition: Identifying a person of interest from a media collection: http://biometrics.cse.msu.edu/Publications/Face/BestRowdenetal.UnconstrainedFaceRecognition.TechReport.MSU-CSE-14-1.pdf

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format. Each image is accompanied by a label indicating the name of the person in the image. While subpopulation data was not available at the initial release of the dataset, a subsequent paper[3] reports the distribution of images by age, race and gender. Table 2 lists these results.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version; c) are there access restrictions or fees?

Everything is included in the dataset.

Are there recommended data splits and evaluation measures? (e.g., training, development, testing; accuracy or AUC)

The dataset comes with specified train/test splits such that none of the people in the training split are in the test split and vice versa. The data is split into two views, View 1 and View 2. View 1 consists of a training subset (pairsDevTrain.txt) with 1100 pairs of matched and 1100 pairs of mismatched images, and a test subset (pairsDevTest.txt) with 500 pairs of matched and mismatched images. Practitioners can train an algorithm on the training set and test on the test set, repeating as often as necessary. Final performance results should be reported on View 2 which consists of 10 subsets of the dataset. View 2 should only be used to test the performance of the final model. We recommend reporting performance on View 2 by using leave-one-out cross validation, performing 10 experiments. That is, in each experiment, 9 subsets should be used as a training set and the 10^{th} subset should be used for testing. At a minimum, we recommend reporting the **estimated mean accuracy,** $\hat{\mu}$ and the **standard error of the mean:** S_E for View 2.

$\hat{\mu}$ is given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \quad (1)$$

where p_i is the percentage of correct classifications on View 2 using subset i for testing. S_E is given as:

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \quad (2)$$

Where $\hat{\sigma}$ is the estimate of the standard deviation, given by:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{10}(p_i - \hat{\mu})^2}{9}} \quad (3)$$

The multiple-view approach is used instead of a traditional train/validation/test split in order to maximize the amount of data available for training and testing.

[3] http://biometrics.cse.msu.edu/Publications/Face/HanJain.UnconstrainedAgeGenderRaceEstimation.MSUTechReport2014.pdf

Training Paradigms: There are two training paradigms that can be used with our dataset. Practitioners should specify the training paradigm they used while reporting results.

- **Image-Restricted Training** This setting prevents the experimenter from using the name associated with each image during training and testing. That is, the only available information is whether or not a pair of images consist of the same person, not who that person is. This means that there would be no simple way of knowing if there are multiple pairs of images in the train/test set that belong to the same person. Such inferences, however, might be made by comparing image similarity/equivalence (rather than comparing names). Thus, to form training pairs of matched and mismatched images for the same person, one can use image equivalence to add images that consist of the same person.

  The files pairsDevTrain.txt and pairsDevTest.txt support image-restricted uses of train/test data. The file pairs.txt in View 2 supports the image-restricted use of training data.

- **Unrestricted Training** In this setting, one can use the names associated with images to form pairs of matched and mismatched images for the same person. The file people.txt in View 2 of the dataset contains subsets of of people along with images for each subset. To use this paradigm, matched and mismatched pairs of images should be formed from images in the same subset. In View 1, the files peopleDevTrain.txt and peopleDevTest.txt can be used to create arbitrary pairs of matched/mismatched images for each person. The unrestricted paradigm should only be used to create training data and not for performance reporting. The test data, which is detailed in the file pairs.txt, should be used to report performance. We recommend that experimenters first use the image-restricted paradigm and move to the unrestricted paradigm if they believe that their algorithm's performance would significantly improve with more training data. While reporting performance, it should be made clear which of these two training paradigms were used for particular test result.

What experiments were initially run on this dataset? Have a summary of those results.

The dataset was originally released without reported experimental results but many experiments have been run on it since then.

Any other comments?

Table 1 summarizes some dataset statistics and Figure 1 shows examples of images. Most images in the dataset are color, a few are black and white.

Property	Value
Database Release Year	2007
Number of Unique Subjects	5649
Number of total images	13,233
Number of individuals with 2 or more images	1680
Number of individuals with single images	4069
Image Size	250 by 250 pixels
Image format	JPEG
Average number of images per person	2.30

Table 1. A summary of dataset statistics extracted from the original paper: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%

Table 2. Demographic characteristics of the LFW dataset as measured by Han, Hu, and Anil K. Jain. *Age, gender and race estimation from unconstrained face images.* Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014).

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API)

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley[4]. The images in this database were gathered from news articles on the web using software to crawl news articles.

Who was involved in the data collection process? (e.g., students, crowdworkers) and how were they compensated (e.g., how much were crowdworkers paid)?

Unknown

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame of the instances?

Unknown

(Gebru et al., 2018)

Data Intelligibility: Datasheets for Datasets

– Questions cover dataset motivation, composition, collection process, pre-processing, distribution, maintenance, legal concerns, and ethical concerns

– Sample use cases:
  - Post with public datasets to inform potential users about the make-up and origin of the data
  - Include with a company’s internal-use datasets to provide relevant information to future users from across the company

No One-Size-Fits-All Solution

	Audit a single prediction	Understand model globally	Make better decisions	Debug models	Assess bias	Inspire trust
CEOs			Approach A			
Data scientists				Approach C		
Lay people						
Regulators	Approach B					

No One-Size-Fits-All Solution

- Why is the explanation needed? What is your goal?
- What is being explained? Prediction or whole system?
- To whom should the system be intelligible?
- Does the explainer have access to system internals?
- Does the explainer have access to the training data?
- What is the dimensionality or scale of the system?
- What type of data is used? Feature vectors? Text?
- Could giving away too much open up the system to manipulation?
- Could giving away too much reveal proprietary information?

Takeaways

- There is no one-size-fits-all solution to fairness, transparency, or intelligibility
- These principles cannot be treated as afterthoughts; they must be considered at every stage of the machine learning pipeline
- Technology can be part of the solution, if used with care
- It is important to involve diverse stakeholders and gather multiple perspectives
- We should admit our mistakes and learn from them

Thanks!

http://jennwv.com

jenn@microsoft.com

Wrapping Up and Looking Ahead: Roundtable Discussion of Key Legal and Regulatory Questions in the Field

Session moderated by:

Ellen Connelly
Federal Trade Commission
Office of Policy Planning

Benjamin Rossen
Federal Trade Commission
Division of Privacy and Identity Protection

Wrapping Up and Looking Ahead: Roundtable Discussion of Key Legal and Regulatory Questions in the Field

Panel Discussion:

Justin Brookman, Pam Dixon,
Salil Mehra, Joshua New,
Nicol Turner-Lee

Moderators: Ellen Connelly & Benjamin Rossen

Closing Remarks

Danielle Holley-Walker
Howard University School of Law

Thank You

Join Us In December