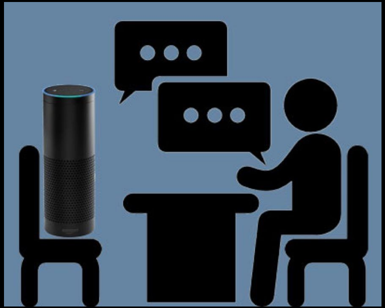
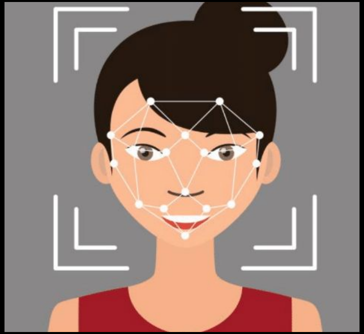
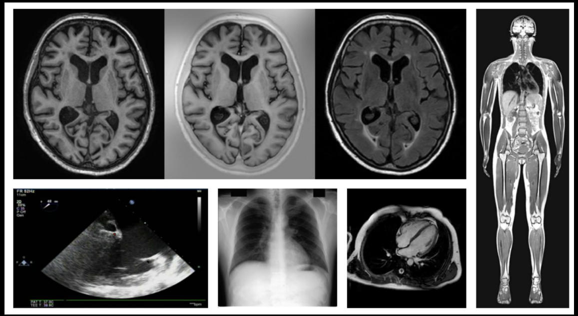
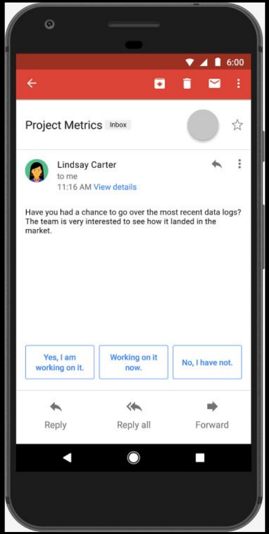
Data, Power, and AI Ethics

Emily Denton
Research Scientist, Google Brain









The man at bat readies to swing at the pitch while the umpire looks on.

“The potential of AI”

“Imagine for a moment that you’re in an office, hard at work.

But it’s **no ordinary office**. By observing cues like your posture, tone of voice, and breathing patterns, it can **sense your mood and tailor the lighting and sound accordingly**. Through gradual ambient shifts, the space around you **can take the edge off when you’re stressed, or boost your creativity when you hit a lull**. Imagine further that you’re a designer, using tools with equally perceptive abilities: at each step in the process, they riff on your ideas based on their knowledge of your own creative persona, contrasted with features from the best work of others.”

[Landay (2019). “Smart Interfaces for Human-Centered AI”]

“The potential of AI”

“Imagine for a moment that you’re in an office, hard at work.

But it’s **no ordinary office**. By observing cues like your posture, tone of voice, and breathing patterns, it can **sense your mood and tailor the lighting and sound accordingly**. Through gradual ambient shifts, the space around you **can take the edge off when you’re stressed, or boost your creativity when you hit a lull**. Imagine further that you’re a designer, using tools with equally perceptive abilities: at each step in the process, they riff on your ideas based on their knowledge of your own creative persona, contrasted with features from the best work of others.”

Potential for who?

[Landay (2019). “Smart Interfaces for Human-Centered AI”]

Another future

“Someday you may have to work in an office where the lights are **carefully programmed and tested by your employer to hack your body**’s natural production of melatonin through the use of blue light, eking out every drop of energy you have while you’re on the clock, leaving you physically and emotionally drained when you leave work. Your eye movements may someday come under the **scrutiny of algorithms** unknown to you that c**lassifies you on dimensions such as “narcissism” and “psychopathy”, determining your career and indeed your life prospects.**”

[Alkhatib (2019). “Anthropological/Artificial Intelligence & the HAI”]

Outline

Part I: Algorithmic (un)fairness

Part II: Data, power, and inequity

Part III: Equitable and accountable AI research

Outline

Part I: Algorithmic (un)fairness

Part II: Data, power, and inequity

Part III: Equitable and accountable AI research

Patterns of exclusion: Object recognition

Object classification accuracy dependent on geographical location and household income

DeVries et al. (2019). [Does Object Recognition Work for Everyone]?

Ground truth: Soap
Nepal, 288 $ / month

Common machine classifications: food, cheese, food product, dish, cooking

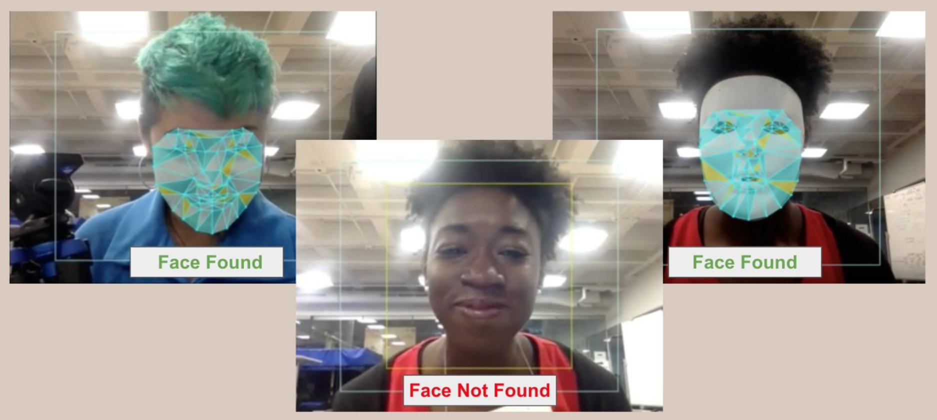
Ground truth: Soap
UK, 1890 $ / month

Common classification: soap dispenser, toiletry, faucet, lotion

Patterns of exclusion: Image classification

[Shankar et al. (2017). No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World]

Patterns of exclusion: Facial analysis

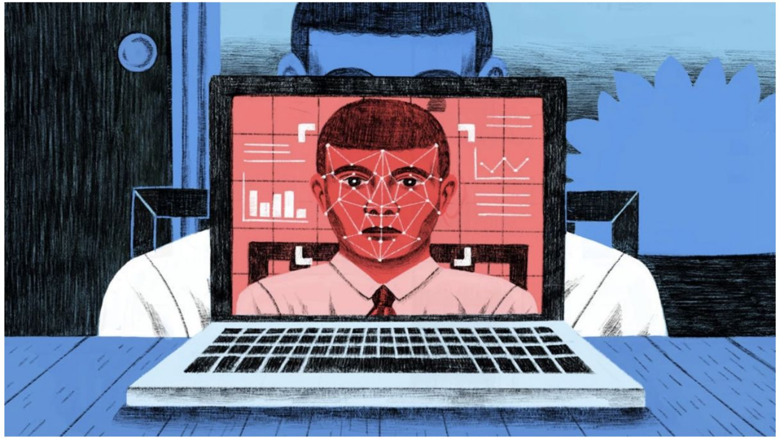


When the Robot Doesn't See Dark Skin

By Joy Buolamwini

Ms. Buolamwini is the founder of the Algorithmic Justice League.

June 21, 2018



"Wearing a white mask worked better than using my actual face" -- Joy Buolamwini

[The Coded Gaze: Unmasking Algorithmic Bias]

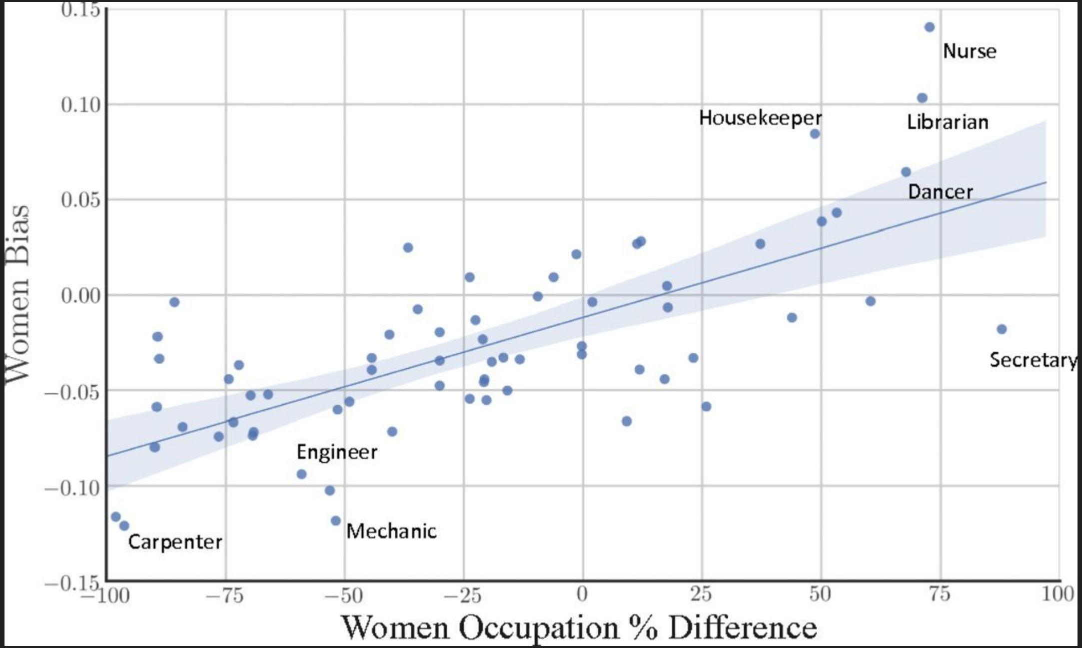
We’ve seen this before…

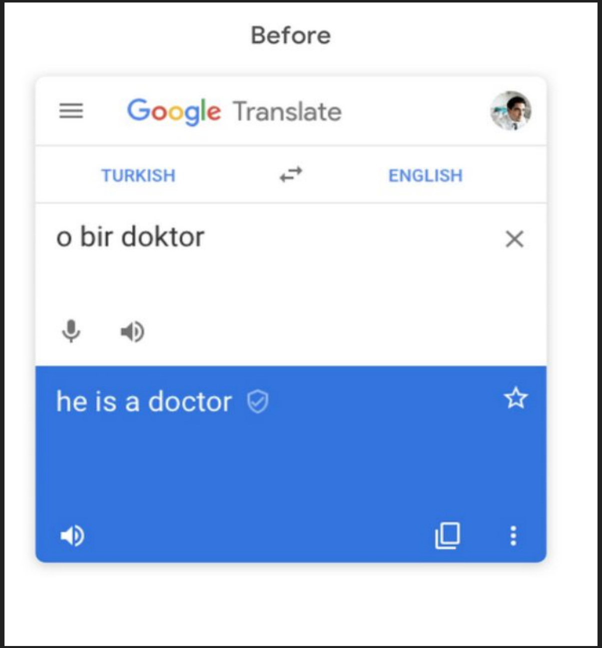
Technology has a long history of encoding whiteness as a default

“Shirley cards” calibrated color film for lighter skin tones

Roth (2009). Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity
Josh Lovejoy (2018). Fair Is Not the Default.

Representational harms: Gender stereotypes in language models



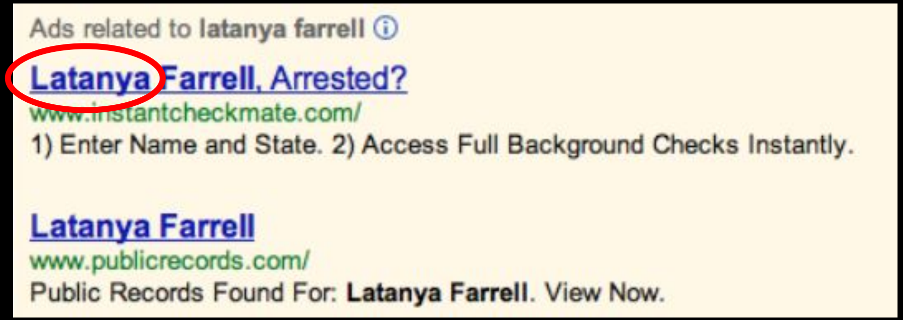


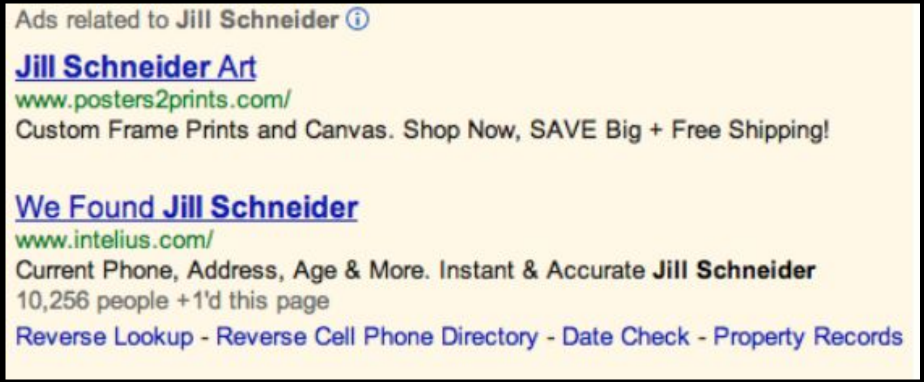
Garg et al. (2018). [Word embeddings quantify 100 years of gender and ethnic stereotypes]

Representational harms: Racial stereotypes in search engines

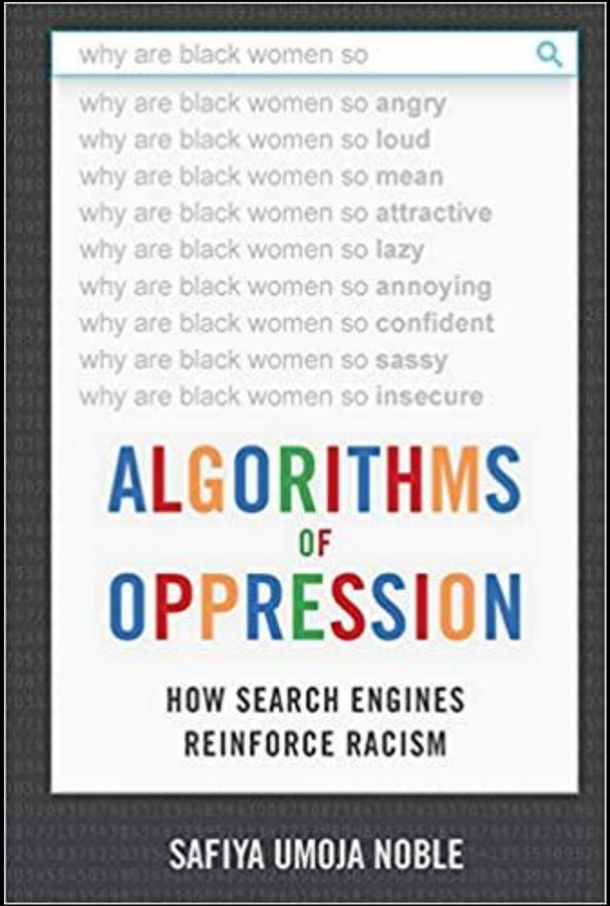
Ads suggestive of arrest record served for queries of Black-associated names

Sweeney (2013). Discrimination in Online Ad Delivery.





Representational harms: Racial stereotypes in search engines



Discrimination in automated decision making tools: Carceral system

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Angwin et al. (2016). Machine Bias.

Discrimination in automated decision making tools: Healthcare

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

NEWS · 24 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

Discrimination in automated decision making tools: Employment

Why Amazon's Automated Hiring Tool Discriminated Against Women

By Rachel Goodman, Staff Attorney, ACLU Racial Justice Program
OCTOBER 12, 2018 | 1:00 PM

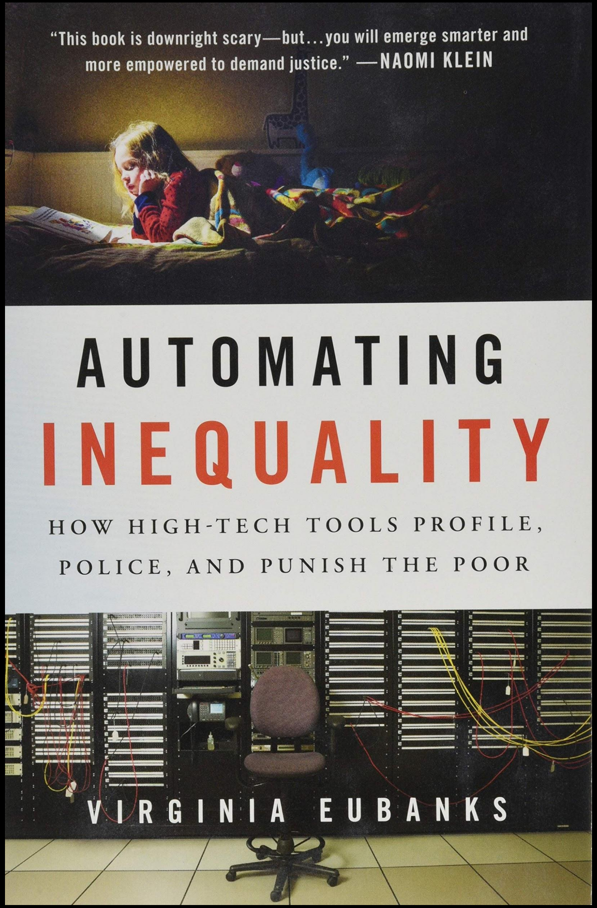
BUSINESS NEWS   OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.

By JORDAN WEISSMANN   OCT 10, 2018 • 4:52 PM

Discrimination in automated decision making tools



AI systems are tools that operate within existing systems of inequality

US ADULTS INDEXED

130 MILLION

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

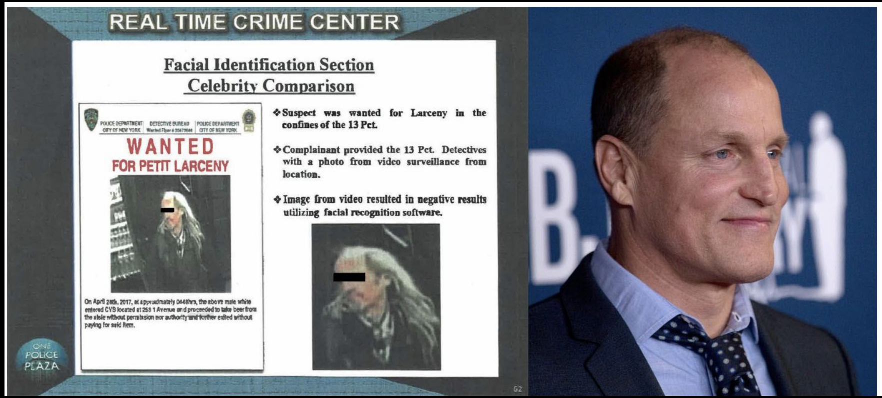
(Garvie , Bedoya, Frankle 2016)

Facial Recognition is the Plutonium of AI

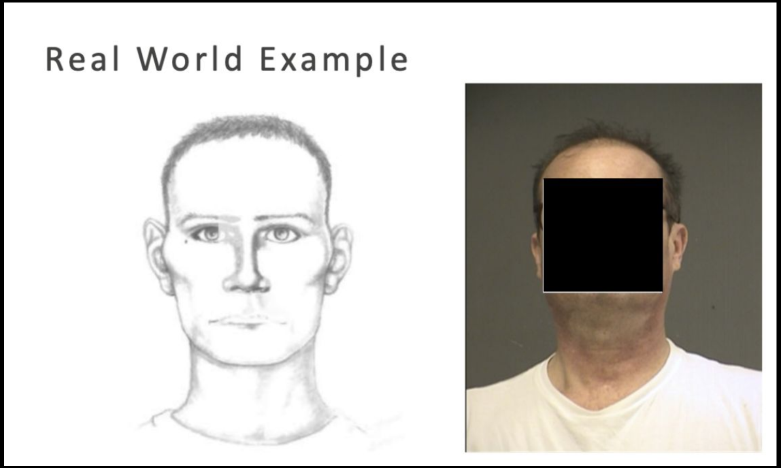
It's dangerous, racializing, and has few legitimate uses; facial recognition needs regulation and control on par with nuclear waste.

By Luke Stark

AI systems are tools that operate within existing systems of inequality



Celebrity faces as probe images



Composite sketches as probe images

[Garvie (2019). Garbage In, Garbage Out: Face Recognition on Flawed Data]

Outline

Part I: Algorithmic (un)fairness

Part II: Data, power, and inequity

Part III: Equitable and accountable AI research

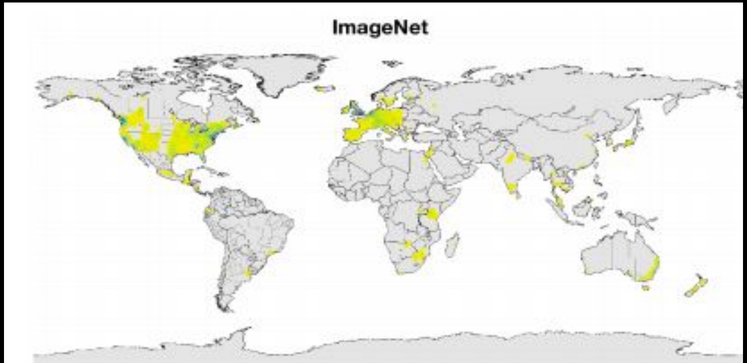
“Every data set involving people implies subjects and objects, those who collect and those who make up the collected. It is imperative to remember that on both sides we have human beings."

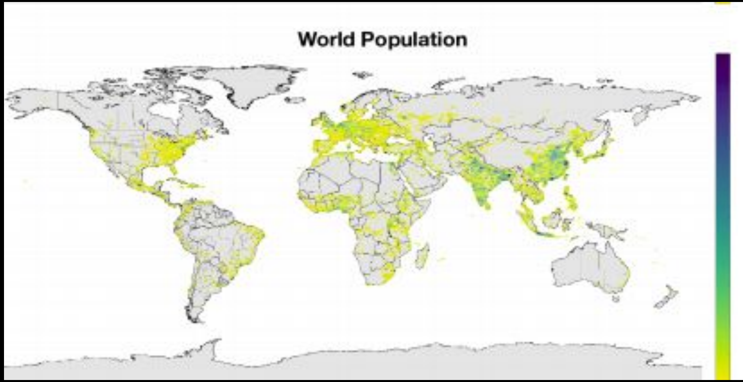
- Mimi Onuoha (2016)

Sampling bias

The selected data is **not representative** of the relevant population

Object recognition datasets





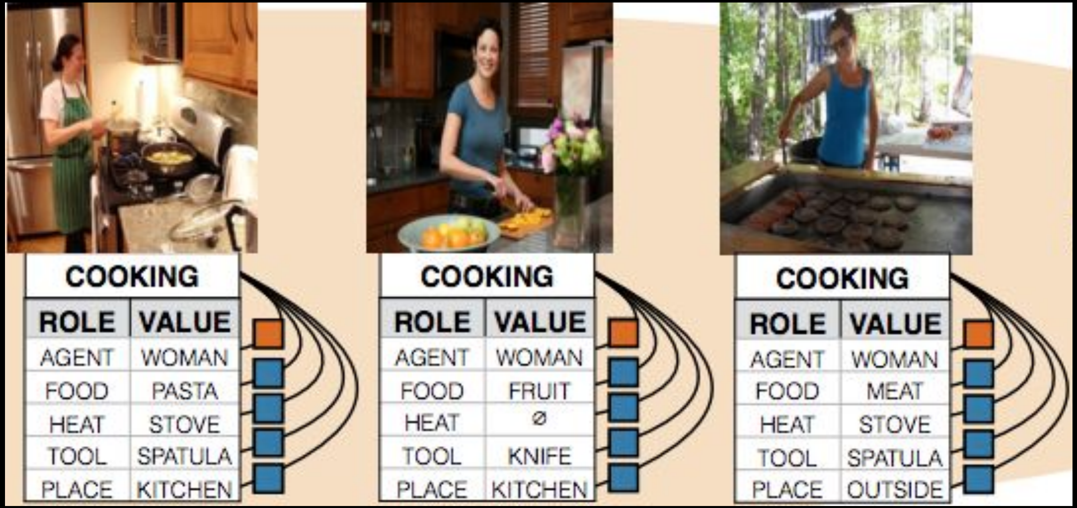
Facial analysis datasets

LFW	77.5% male 83.5% white
IJB-A	79.6% lighter-skinned
Adience	86.2% lighter-skinned

Buolamwini & Gebru (2018). [Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification]
DeVries et al. (2019). [Does Object Recognition Work for Everyone?]

Sampling bias

Approx 50% of verbs in imSitu visual semantic role labeling (vSRL) dataset are extremely biased in the male or female direction



`shopping`, `cooking` and `washing` biased towards women
`driving`, `shooting`, and `coaching` biased towards men

[Zhao et al. (2017) [Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints]]

Human reporting bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals.

Reporting bias

World learning from text

Word	Frequency in corpus
"spoke"	11,577,917
"laughed"	3,904,519
"murdered"	2,834,529
"inhaled"	984,613
"breathed"	725,034
"hugged"	610,040
"blinked"	390,692
"was late"	368,922
"exhaled"	168,985
"was punctual"	5,045

Gordon and Van Durme (2013). Reporting Bias and Knowledge Acquisition

Reporting bias

World learning from text

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“was late”	368,922
“exhaled”	168,985
“was punctual”	5,045

Gordon and Van Durme (2013). Reporting Bias and Knowledge Acquisition

Reporting bias

What do you see?

“Bananas”

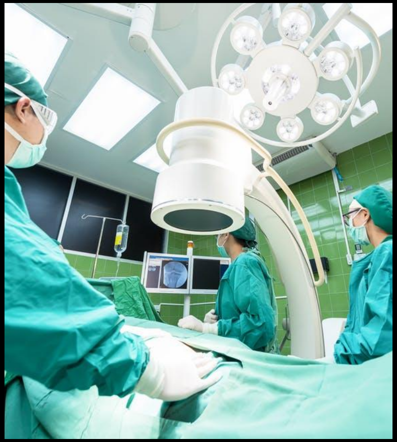
“**Green** bananas”
“Unripe bananas”

[Misra et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels]

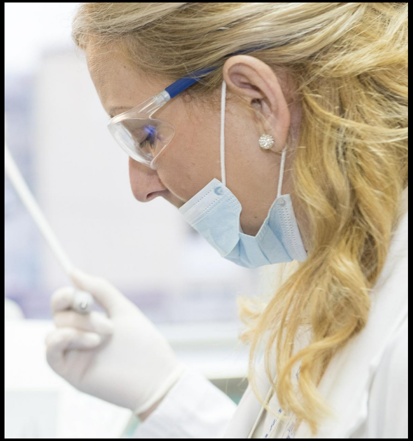
Reporting bias

Social stereotypes can affect implicit prototypicality judgements

“Doctor”



“**Female** doctor”



Implicit stereotypes

Unconscious attribution of characteristics, traits and behaviours to members of certain social groups.

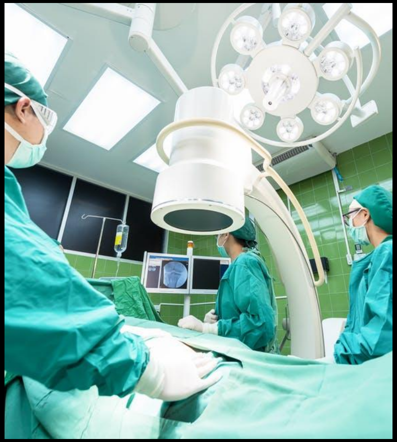
Data annotation tasks can activate implicit social stereotypes.

Implicit gender stereotypes

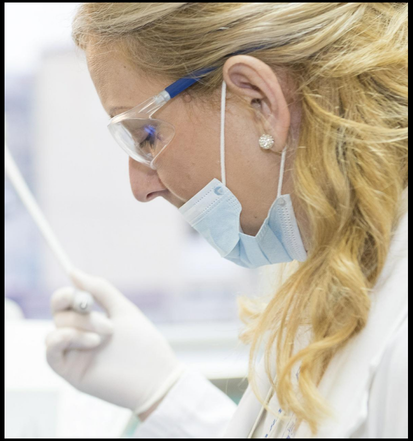
Implicit biases can also affect how people classify images

Filter into a computer vision system through annotations

“Doctor”



“Nurse”



Historical bias

Biases that arise from the world as it was when the data was sampled.

Historical bias

If historical hiring practices favor men, gendered cues in the data will be predictive of a ‘successful candidate’

Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.

By JORDAN WEISSMANN

OCT 10, 2018 • 4:52 PM

~~Historical bias~~

Historical (and ongoing) injustices encoded in datasets

~~Historical bias~~

Historical (and ongoing) injustices encoded in datasets

Systemic racism and sexism is *foundational* all our major institutions

Data is generated through social processes and reflects the social world

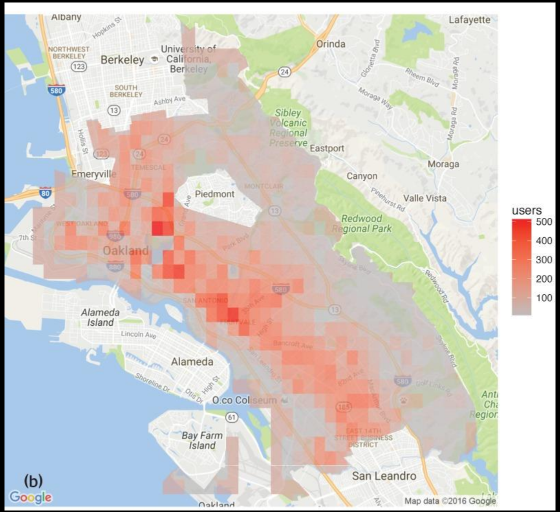
‘Unbiased’ data is a myth that obscures the entanglement between tech development and structural inequality

Policing and surveillance applications

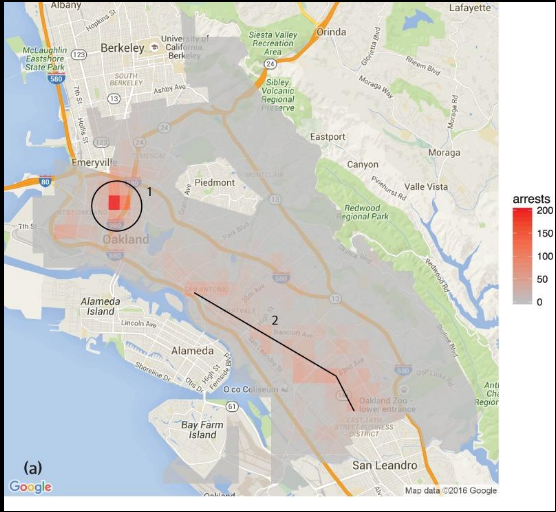
Predictive policing tools predict "crime hotspots" based on policing data that reflects corrupt and racially discriminatory practices of policing and documentation

Lum & Isaac (2016). To predict and serve?

Richardson et al. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice



Estimated number of drug users, based National Survey on Drug Use and Health



Drug arrests made by Oakland police department

“When bias is routed through technoscience and coded ‘scientific’ and ‘objective’ … it becomes even more difficult to challenge it and hold individuals and institutions accountable.”

- Ruha Benjamin, *Race After Technology*

Policing and surveillance applications: Who defines ‘high risk’?

Clifton et al. (2017). *White Collar Crime Risk Zones*

Healthcare applications

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

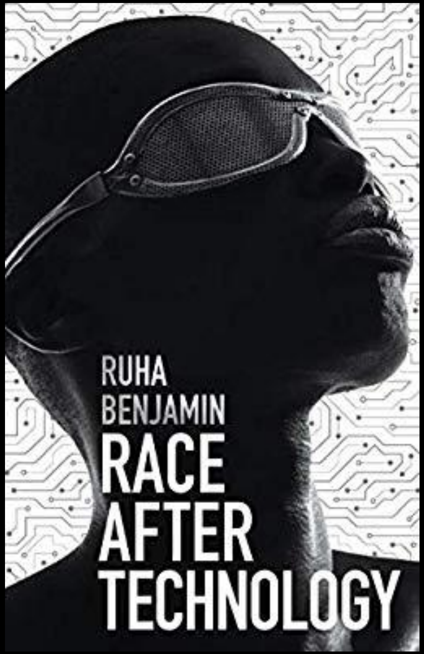
+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

NEWS · 24 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

“New Jim Code": ‘race neural’ algorithms that reproduce racial inequality



Datasets construct a particular view of the world -- a view that is often laden with subjective values, judgements, & imperatives

Data is always always socially and culturally situated (Gitelman, 2013; Elish and boyd, 2017)

Datasets construct a particular view of the world -- a view that is often laden with subjective values, judgements, & imperatives

This is inescapable

There is no “view from nowhere” (Haraway, 1991)

The view of the world through ImageNet

"To produce a dataset at 'the scale of the web' implies to impose a particular way of seeing images, of pointing and naming." -- Malevé (2019)

Hammerhead shark → Scientific object

Trout → Dead trophy

Lobster → Food

The view of the world through ImageNet

The women of ImageNet ➔ Bikinis and mini-skirts

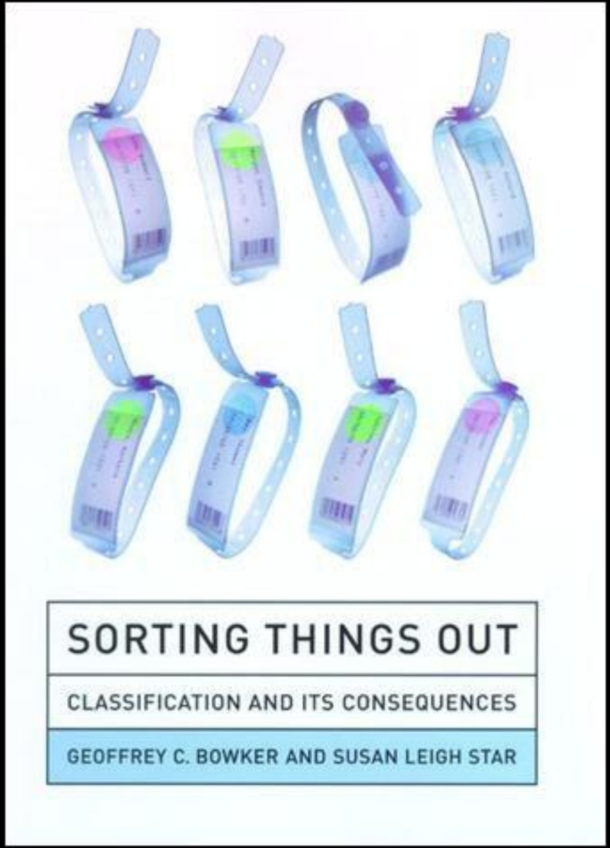
The men of ImageNet ➔ Music, sports, and fishing

Prabhu & Birhane (2020). Large image datasets: A pyrrhic win for computer vision?

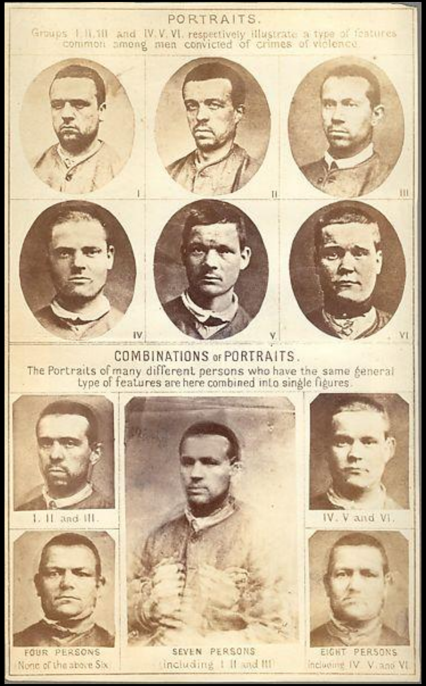
The politics of classification

Classifications within within machine learning datasets reflect sociotechnical decisions and embed politics, values, and power imbalances

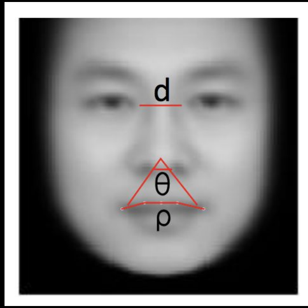
Data-driven doesn't inherently imply empirically grounded and scientific

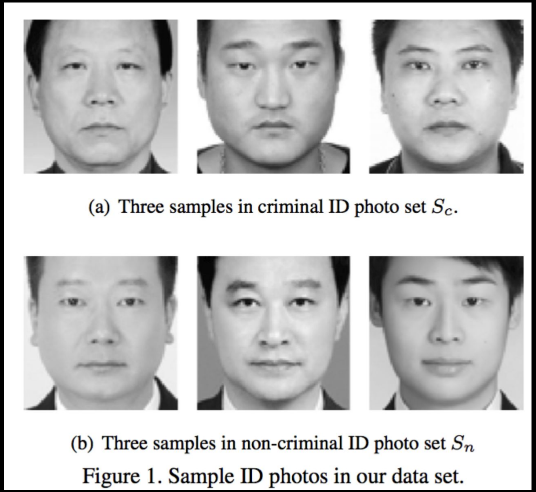


Technologies of human classification



Francis Galton (1877). Composite portraits of human 'types'





Wu and Zhang (2016). Automated Inference on Criminality using Face Images

Technologies of human classification

Aguera y Arcas (2017). Physiognomy's New Clothes

HOW TO TELL JAPS FROM THE CHINESE

ANGRY CITIZENS VICTIMIZE ALLIES WITH EMOTIONAL OUTBURST AT ENEMY

In the first discharge of emotions touched off by the Japanese assaults on their nation, U. S. citizens have been demonstrating a distressing ignorance on the delicate question of how to tell a Chinese from a Jap. Innocent victims in cities all over the country are many of the 75,000 U. S. Chinese, whose homeland is our stanch ally. So serious were the consequences threatened, that the Chinese consulates last week prepared to tag their nationals with identification buttons. To dispel some of this confusion, LIFE here adduces a rule-of-thumb from the anthropometric conformations that distinguish friendly Chinese from enemy alien Japs.

To physical anthropologists, devoted debunkers of race myths, the difference between Chinese and Japs is measurable in millimeters. Both are related to the Eskimo and North American Indian. The modern Jap is the descendant of Mongoloids who invaded the Japanese archipelago back in the mists of prehistory, and of the native aborigines who possessed the islands before them. Physical anthropology, in consequence, finds Japs and Chinese as closely related as Germans and English. It can, however, set apart the special types of each national group.

The typical Northern Chinese, represented by Ong Wen-hao, Chungking's Minister of Economic Affairs (*left, above*), is relatively tall and slenderly built. His complexion is parchment yellow, his face long and delicately boned, his nose more finely bridged. Representative of the Japanese people as a whole is Premier and General Hideki Tojo (*left, below*), who betrays aboriginal antecedents in a squat, long-torsoed build, a broader, more massively boned head and face, flat, often pug, nose, yellow-ocher skin and heavier beard. From this average type, aristocratic Japs, who claim kinship to the Imperial Household, diverge sharply. They are proud to approximate the patrician lines of the Northern Chinese.

…hinese public servant, Ong Wen-hao, is representative of …rth Chinese anthropological group with long, fine-boned … and scant beard. Epicanthic fold of skin above eyelid …found in 85% of Chinese. Southern Chinese have round, broad faces, not as massively boned as the Japanese. Except that their skin is darker, this description fits Filipinos who are often mistaken for Japs. Chinese sometimes pass for Europeans; but Japs more often approach Western types.

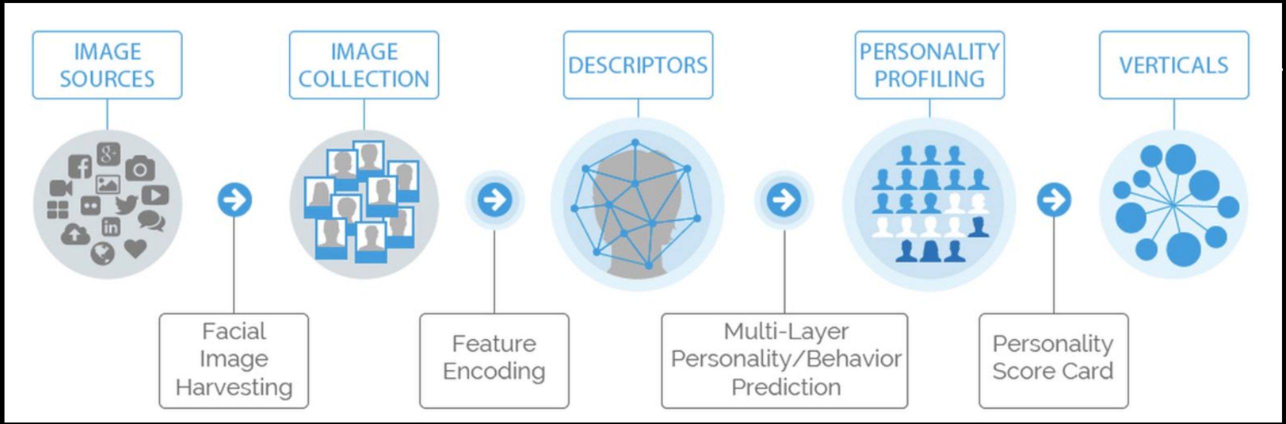
…panese warrior, General Hideki Tojo, current Premier, is …murai, closer to type of humble Jap than highbred rela… of Imperial Household. Typical are his heavy beard, …ssive cheek and jaw bones. Peasant Jap is squat Mongoloid, with flat, blob nose. An often sounder clue is facial expression, shaped by cultural, not anthropological, factors. Chinese wear rational calm of tolerant realists. Japs, like General Tojo, show humorless intensity of ruthless mystics.

Chinese journalist, Joe Chiang, found it necessary to advertise his nationality to gain admittance to White House press conference. Under Immigration Act of 1924, Japs and Chinese, as members of the "yellow race," are barred from immigration and naturalization.

Jo & Gebru (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning

“Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for **profiling people** and revealing their personality **based only on their facial image.”**

- Faception startup



“High IQ”

“White-Collar Offender”

“Terrorist”

Datasets represent specific formulations of a problem

Fairness concerns often stem from decisions about how to operationalize social constructs within a datasets (Jacobs and Wallach, 2018)

Crime patterns	↔	Policing patterns
Illness	↔	Health care costs
Successful job candidate	↔	Hiring and retention patterns

Outline

Part I: Algorithmic (un)fairness

Part II: Data, power, and inequity

Part III: Equitable and accountable AI research

Ethics-informed model testing

Consider **multiple evaluation metrics** - they each provide different information

		Model Predictions	
		Positive ($\hat{Y}$= 1)	**Negative** ($\hat{Y}$ = 0)
Target	**Positive** (Y= 1)	True positives	False negatives
	Negative (Y= 0)	False negatives	True negatives

Ethics-informed model testing

Consider **multiple evaluation metrics** - they each provide different information

Compute metrics over subgroups defined along cultural, demographic, phenotypical lines

❖ How you define groups will be context specific

Evaluate for each (metric, subgroup) pair

Ethics-informed model testing

Unitary groups

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

[Buolamwini and Gebru, 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification]

Ethics-informed model testing

Intersectional groups

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

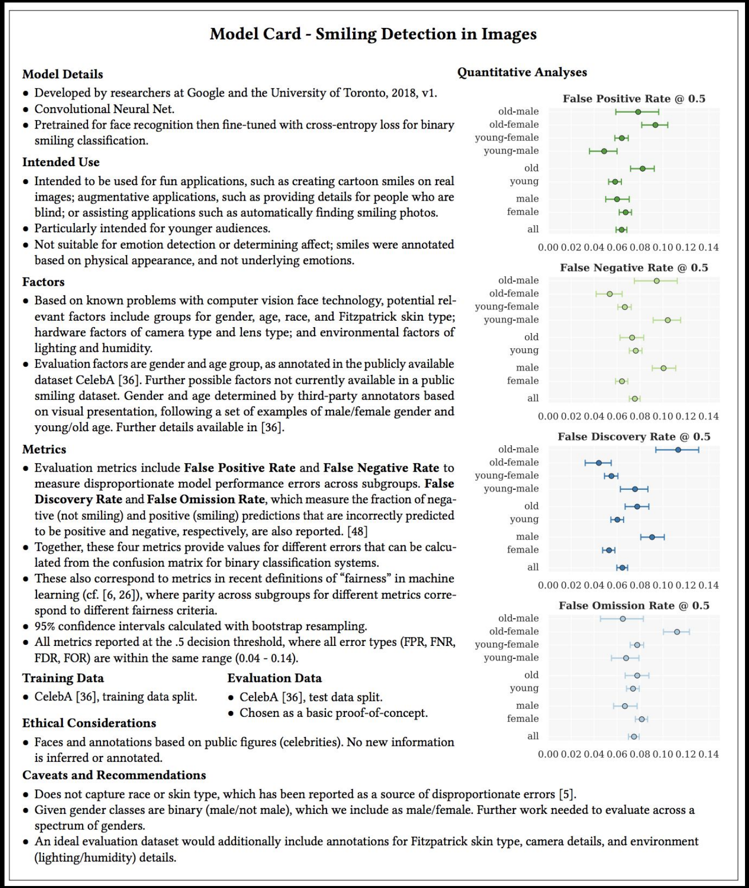
[Buolamwini and Gebru, 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification]

Model and data transparency

Model cards: Standardized framework for transparent model reporting

Model creators:
Encourage thorough and critical evaluations
Outline potential risks or harms, and implications of use

Model consumers:
Provide information to facilitate informed decision making



Model Card - Smiling Detection in Images

Model Details
- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use
- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors
- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data
- CelebA [36], training data split.

Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses

Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Mitchell et al. (2019). [Model Cards for Model Reporting]

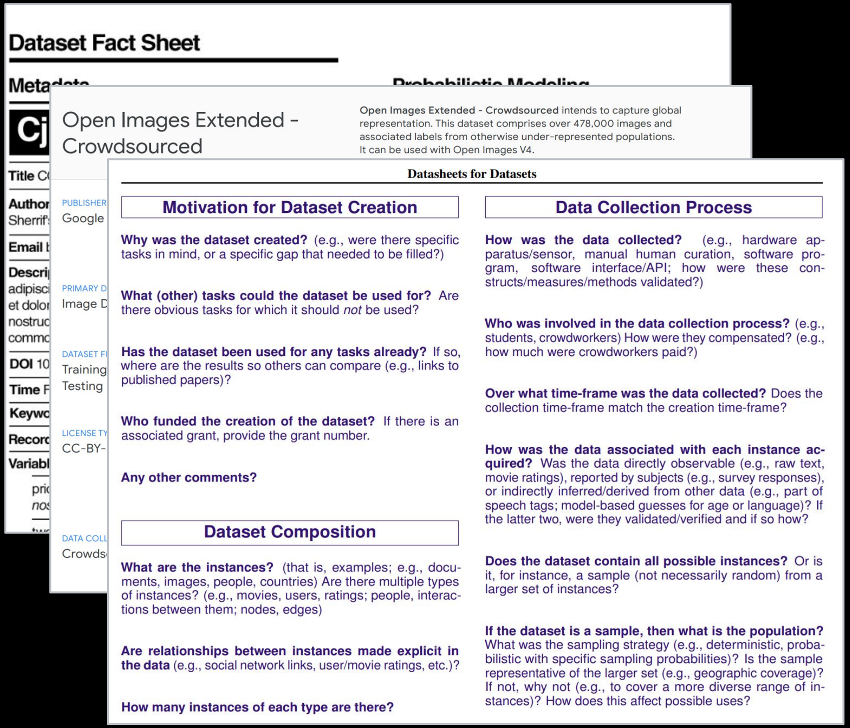
Model and data transparency

Standardized framework for transparent dataset documentation

Dataset creators:

Reflect on on process of creation, distribution, and maintenance
Making explicit any underlying assumptions
Outline potential risks or harms, and implications of use

Dataset consumers:
Provide information to facilitate informed decision making



Timnit, et al. (2018). Datasheets for datasets
Holland et al. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards
Bender and Friedman (2018). Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science

Measurement and construct validity

Fairness concerns often stem from decisions about how to operationalize social constructs within a datasets (Jacobs and Wallach, 2018)

Crime patterns	$\leftrightarrow$	Policing patterns
Illness	$\leftrightarrow$	Health care costs
Successful job candidate	$\leftrightarrow$	Hiring and retention patterns

As a field, we need to rethink how we develop and use datasets

Currently:

- Data decisions go heavily undocumented (Geiger et al. 2020; Scheuerman et al. 2020)

As a field, we need to rethink how we develop and use datasets

Currently:

- Data decisions go heavily undocumented (Geiger et al. 2020; Scheuerman et al. 2020)

- Categories tend to be presented as natural
  - Even highly political categories such as race and gender tend to be presented as indisputable and natural (Scheuerman et al. 2020)

As a field, we need to rethink how we develop and use datasets

Currently:

- Data decisions go heavily undocumented (Geiger et al. 2020; Scheuerman et al. 2020)

- Categories tend to be presented as natural
    - Even highly political categories such as race and gender tend to be presented as indisputable and natural (Scheuerman et al. 2020)

- Annotation and labelling is rarely viewed as interpretive work (Miceli et al. 2020)
    - Annotation demographics often underspecified -- annotators presumed interchangeable

As a field, we need to rethink how we develop and use datasets

Currently:

- Data decisions go heavily undocumented (Geiger et al. 2020; Scheuerman et al. 2020)

- Categories tend to be presented as natural
  - Even highly political categories such as race and gender tend to be presented as indisputable and natural (Scheuerman et al. 2020)

- Annotation and labelling is rarely viewed as interpretive work (Miceli et al. 2020)
  - Annotation demographics often underspecified -- annotators presumed interchangeable

- Ground truth often presumed to be fact (Aroyo & Welty, 2015; Muller et al. 2019)

As a field, we need to rethink how we develop and use datasets

Currently:

- Data work is heavily undervalued, relative to model work
    - NLP dataset publications devalued within peer-review processes (Heinzerling, 2019); ongoing work indicates similar pattern in computer vision

As a field, we need to rethink how we develop and use datasets

Currently:

- Data work is heavily undervalued, relative to model work
    - NLP dataset publications devalued within peer-review processes (Heinzerling, 2019); ongoing work indicates similar pattern in computer vision

- ML curriculums and textbooks don't treat dataset development as a specialty
    - Jo & Gebru, 2020 characterize resulting practices by a *laissez faire* attitude

As a field, we need to rethink how we develop and use datasets

Contingent → Datasets are contingent on the social conditions of creation

Constructed → Data is not objective; ‘Ground truth’ isn’t truth

Value-laden → Datasets are shaped by patterns of inclusion and exclusion

---

Our data collection and data use practices should reflect this

Data is contingent, constructed, value-laden

Who is reflected in the data?

What taxonomies are imposed?

How are images categorized?

Who is doing the categorization?

CelebA dataset

AI research is not a value-neutral endeavor

~~“I’m just an engineer”~~

~~“I’m just doing basic research”~~

Data Science as Political Action

Grounding Data Science in a Politics of Justice

Ben Green
bgreen@g.harvard.edu
Berkman Klein Center for Internet & Society at Harvard University
Harvard John A. Paulson School of Engineering and Applied Sciences

Accountability for the intended and unintended impacts of our work

Status quo is the default, but the status quo is political

"*Detachment in the face of history ensures its ongoing codification*" -- Ruha Benjamin

Shift focus from *intent* ➔ *impact*

Research is contingent and situated -- be attentive to your own positionality

Our social positions in the world and set of experiences shapes and bounds our view of the world; this in turn affects the research questions we pursue and how we pursue them

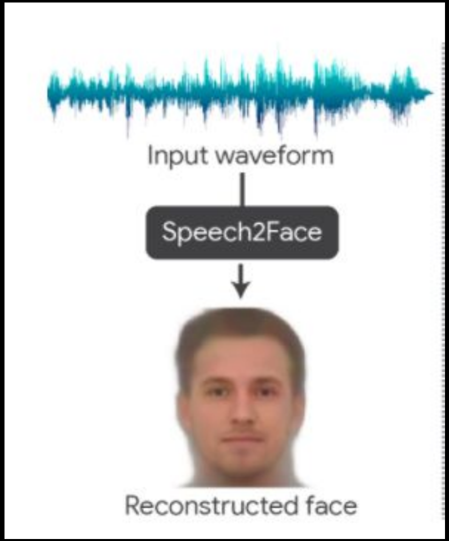
Suggested readings:

Harding (1993). Rethinking Standpoint Epistemology: What is "Strong Objectivity?

Kaeser-Chen et al. (2020). Positionality-Aware Machine Learning

Research is contingent and situated -- be attentive to your own positionality

Limits in your knowledge don't absolve you of responsibility



Voice-to-face synthesis:

Fun application of conditional generative models?

Assistive technology?

Surveillance technology?

Trans-exclusionary technology?

Oh, et al. (2019). [Speech2Face: Learning the Face Behind a Voice]
Wen et al. (2019). [Reconstructing faces from voices]

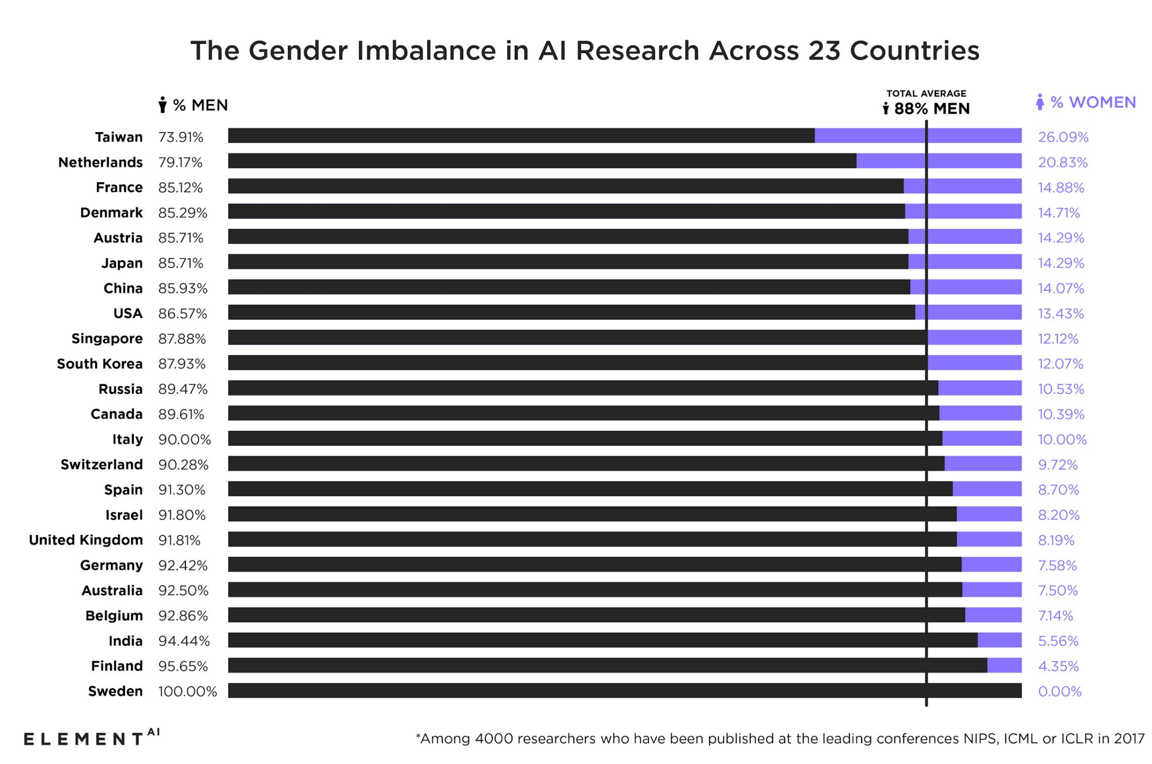
Value knowledge and experience of individuals holding marginalized identities

AI development cannot be divorced from the larger social and political landscape

Who gets a say in the development of AI? Who is most likely to experience positive benefit of AI technologies?

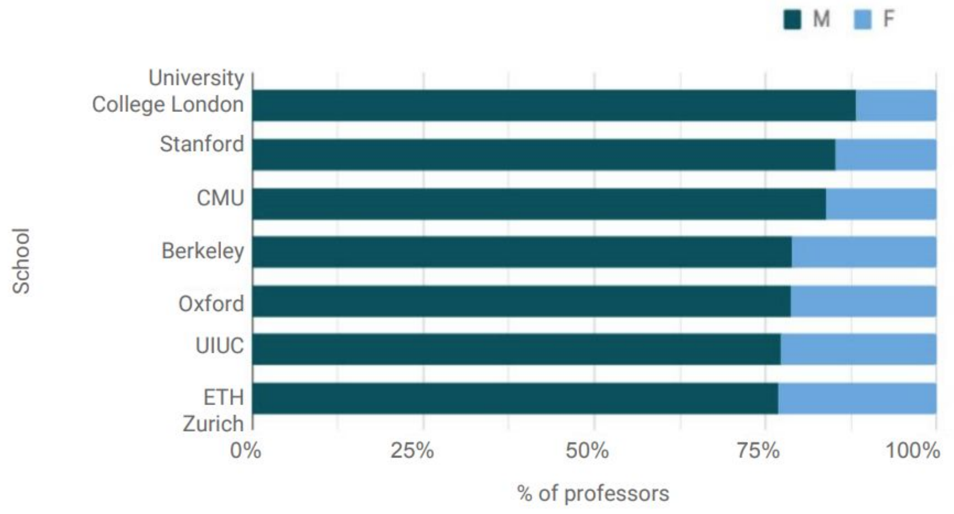
Who is marginalized from AI development? Who is most likely to be harmed by AI technologies?

Diversity and inclusion efforts are part and parcels of responsible AI development



Suggested reading:

West et al. (2019). Discriminating Systems: Gender, Race and Power in AI



80% of AI professors are male

On average, 80% of professors from UC Berkeley, Stanford, UIUC, CMU, UC London, Oxford, and ETH Zurich are male

Facebook (as of 2018)

- ❖ 22% of technical roles filled by women
- ❖ 15% of AI researchers were women

Google (as of 2018)

- ❖ 21% of technical roles filled by women
- ❖ 10% of AI researchers were women

No reported data on trans and non-binary employees, or other gender minorities

Tom Simonite (2018). AI Is the Future—But Where Are the Women?

Facebook (as of 2018)

- 4% Black workers
- 5% Hispanic workers

Microsoft (as of 2018)

- 4% Black workers
- 6% Latinx workers

Google (as of 2018)

- 2.5% Black workers
- 3.6% Latinx workers

West et al. (2019). Discriminating Systems: Gender, Race and Power in AI

Minority tax

Fixing D&I problems

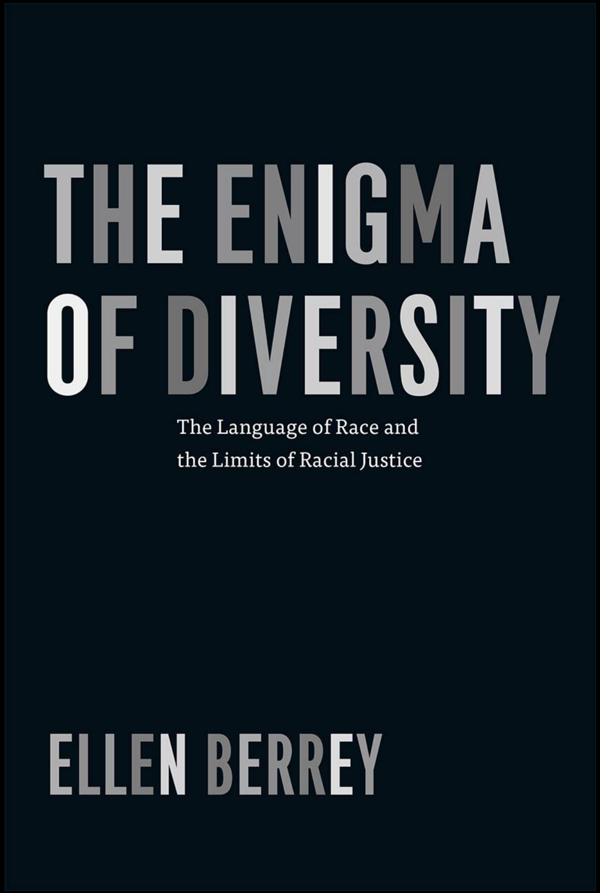
Calling out unethical practices

Interrogate how structural racism, sexism, etc. shape academic and industry hiring practices, cultures, and incentive structures


DISCRIMINATING SYSTEMS
Gender, Race, and Power in AI

Sarah Myers West, AI Now Institute, New York University
Meredith Whittaker, AI Now Institute, New York University, Google Open Research
Kate Crawford, AI Now Institute, New York University, Microsoft Research

APRIL 2019



Value interdisciplinarity and ‘non-technical’ work

Building AI is simultaneously a technical and social endeavour

Racial literacy is important for every AI developer (see Data and Society’s Advancing Racial Literacy in Tech)

Knowledge hierarchies embedded within STEM structure the types of knowledge that is seen as valuable

Lived experiences of individuals experiencing the harms of AI technologies is a form of valuable knowledge

Value knowledge and experience of individuals holding marginalized identities

Those belonging to marginalized groups experience the world in ways that give them access to knowledge that those with the dominant perspective do not

Suggested reading:

Donna Haraway(1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective

Patricia Hill Collins (1990). Black Feminist Thought: Knowledge, Consciousness and the Politics of Empowerment

Sandra Harding (1991). Whose Science? Whose Knowledge?: Thinking from Women's Lives

Value knowledge and experience of individuals holding marginalized identities

Actively follow the perspectives of people in marginalized groups

Listen to your colleagues who have personal experiences with the harms of AI systems

Use your voice and position of power to amplify the voices of marginalized individuals

Learn about design frameworks and organizations that are privilege the perspectives of marginalized stakeholders and are leveraging data to empower marginalized communities (e.g. Design Justice Network, Our Data Bodies, Data for Black Lives)

Thanks!

Emily Denton
dentone@google.com
@cephaloponderer