Fairness Accountability Transparency and Ethics in Computer Vision

Timnit Gebru
Emily Denton

Survey responses, discuss...

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"]

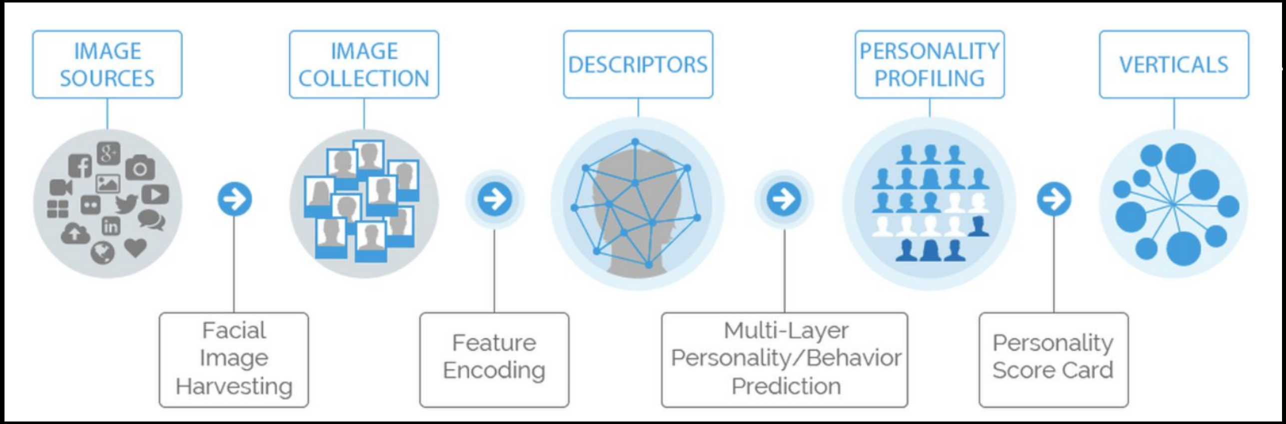
Potential for *who*?

“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 classifies you on dimensions such as “narcissism” and “psychopathy”, determining your career and indeed your life prospects.”

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

"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.**"

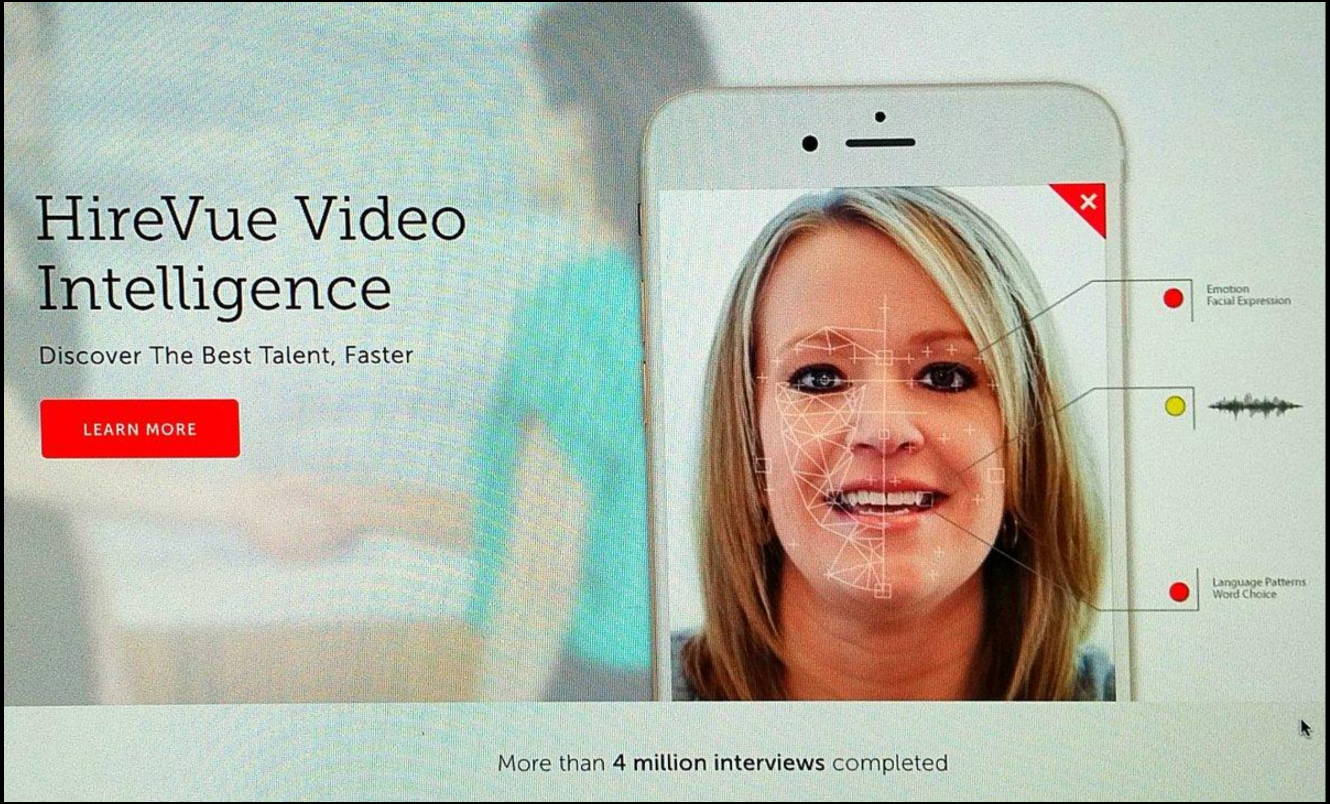
- Faception startup



"High IQ"

"White-Collar Offender"

"Terrorist"


HireVue Video Intelligence
Discover The Best Talent, Faster
LEARN MORE
Emotion
Facial Expression
Language Patterns
Word Choice
More than 4 million interviews completed

“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, Data & Society

Our data bodies

https://www.odbproject.org/

Why We’re Concerned About Data

“Data-based technologies are changing our lives, and the systems our communities currently rely on are being revamped. These data systems do and will continue to have a profound impact on our ability to thrive.

To confront this change, we must first understand how we are both hurt and helped by data-based technologies. ***This work is important because our data is our stories. When our data is manipulated, distorted, stolen, or misused, our communities are stifled, and our ability to prosper decreases.***”

Seeta Pena Gangadharan: A Filipino-Indian mother and research justice organizer, born in New Jersey and teaching in London.

Excerpts from Keynote at Towards Trustworthy ML: Rethinking Security and Privacy for ML ICLR 2020

“People are caught in a never ending cycle of disadvantage based on data that was collected on them. Jill: I plead guilty to worthless checks in 2003: 15 years ago. But this is still being held against me. All of my jobs have been temporary positions.”

“**Refusal.** People refused to settle for the data driven systems: process of data collection systems that were handed to them. Mellow fought tooth and nail to find housing. Repeatedly denied housing. Had witnessed the death of a friend. Each time she re-applied for housing, she was denied….She challenged the data used to categorize her.”

“Ken, a native american man, he deliberately misrepresented himself....The police issued him a ticket without a surname...Ken was practicing refusal against database dependent police practices.”

“**The Problem with Abstraction.** *I have heard computer scientists present their research in relation to real world problems: as if computer scientists and their research is not done in the real world. I listened to papers that tended to disappear people into mathematical equations…”*

"Marginalized people are demonized, deprived. What is the point of making data driven systems 'fairer' if they're going to make institutions colder and more punitive?"

Who is seen? *How* are they seen?

Error Rate$_{(1\text{-}PPV)}$ By Female x Skin Type

	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%

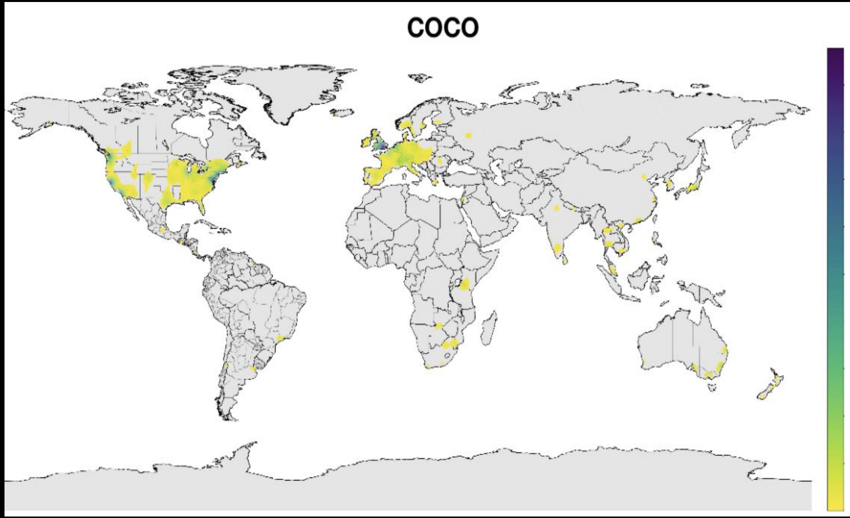
Buolamwini & Gebru FAT* 2018, Slides from Joy Buolamwini

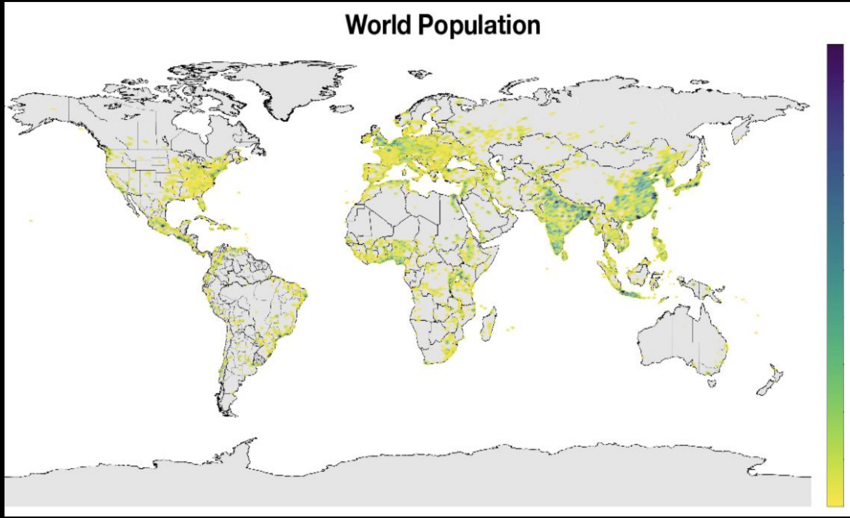
Dataset bias

LFW [Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Huang et al.]	77.5% male 83.5% white
IJB-A [Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark. Klare et al.]	79.6% lighter-skinned
Adience [Age and gender classification using convolutional neural networks. Levi and Hassner.]	86.2% lighter-skinned

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

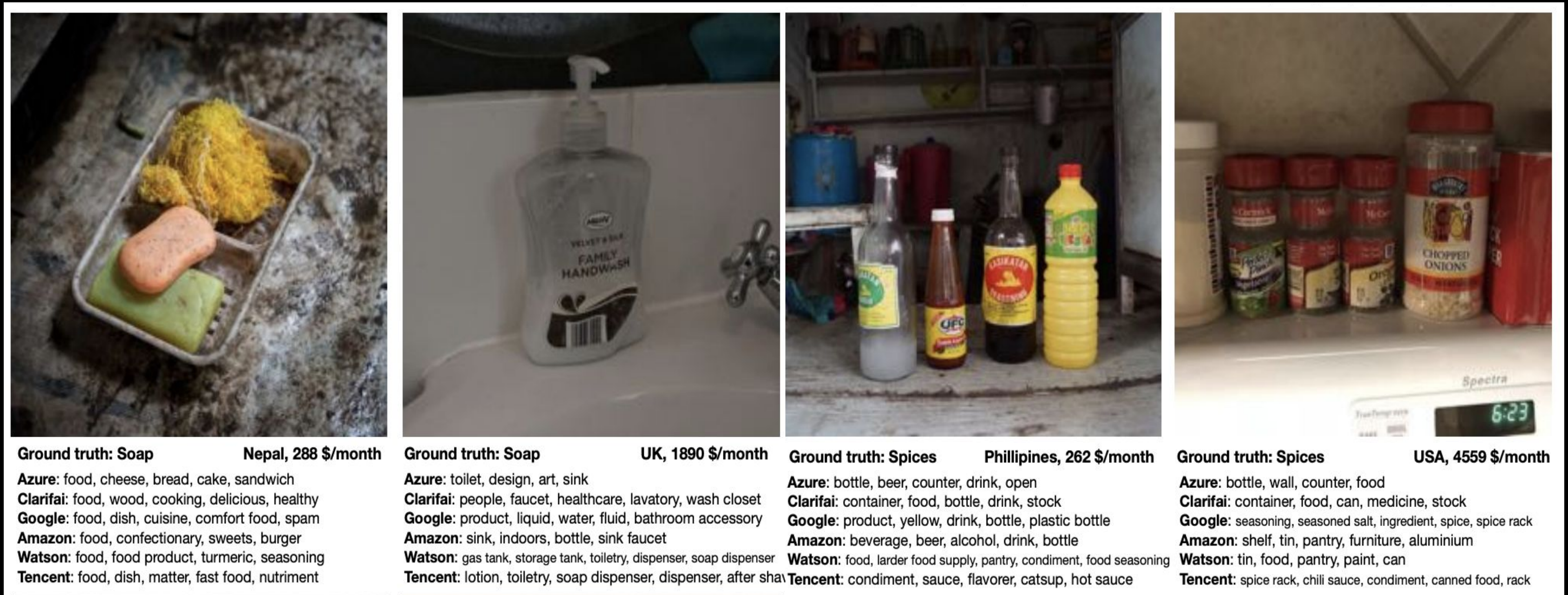
Who is seen? *How* are they seen?





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

Who is seen? *How* are they seen?



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

Who is seen? *How* are they seen?

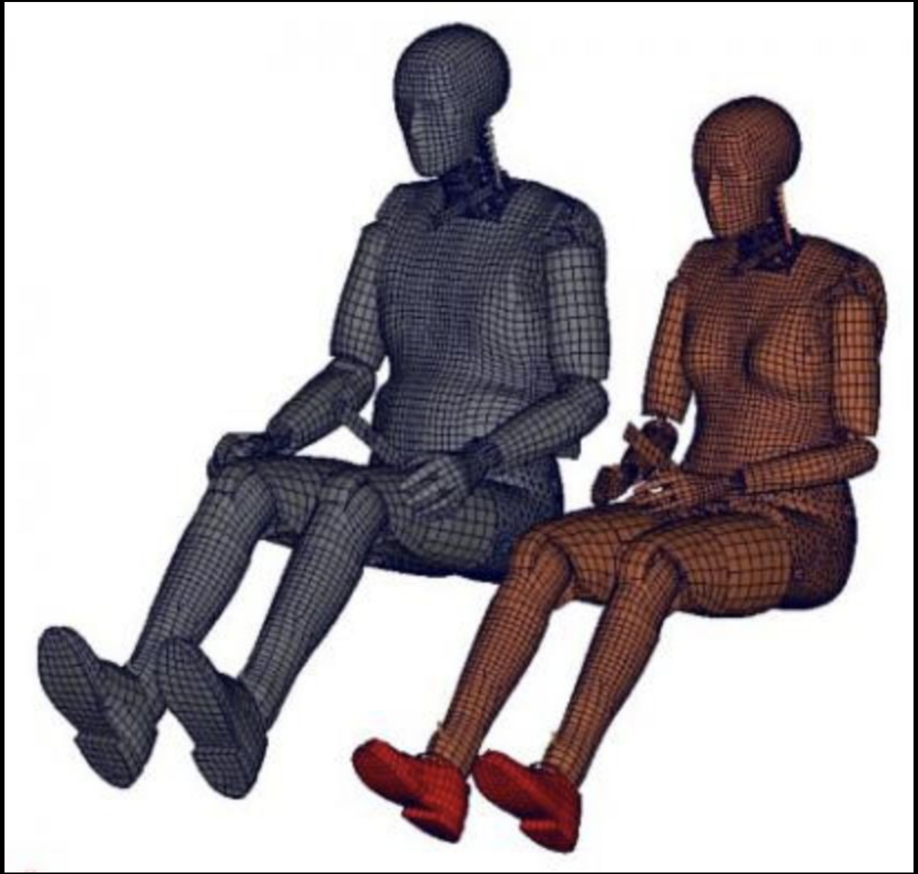
[Shankar et al. (2017). No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World]

Not unique to AI...

The deadly truth about a world built for men – from stab vests to car crashes

Crash-test dummies based on the 'average' male are just one example of design that forgets about women – and puts lives at risk

It wasn't until 2011 that the US started using a female crash-test dummy. Photograph: Kellie French/The Guardian



Not unique to AI...

Your Drugs Probably Weren't Tested on People of Color

Medical research doesn't reflect America's diversity, and it's created a health and economic disaster.

The medical research gender gap: how excluding women from clinical trials is hurting our health

Large gender gaps in research limit how much we know about the difference between women's health and men's

- **This article is part of a series on women's health and chemicals**
- **Women still do most of the cleaning: is it putting their health at risk?**
- **How excluding women from clinical trials is hurting our health**

Visibility is not inclusion

We can’t ignore social & structural problems

Diversity in Faces Dataset

The Diversity in Faces(DiF)is a large and diverse dataset that seeks to advance the study of fairness and accuracy in facial recognition technology.The first of its kind available to the global research community,DiF provides a dataset of annotations of 1 million human facial images.

Access dataset

Read the research paper

Microsoft improves facial recognition technology to perform well across all skin tones, genders

June 26, 2018 | John Roach

A person takes part in a Google facial recognition project. (Obtained by Daily News)

Google using dubious tactics to target people with 'darker skin' in facial recognition project: sources

By GINGER ADAMS OTIS and NANCY DILLON
NEW YORK DAILY NEWS | OCT 02, 2019 | 6:56 PM

Transgender YouTubers had their videos grabbed to train facial recognition software

In the race to train AI, researchers are taking data first and asking questions later

By James Vincent | Aug 22, 2017, 10:44am EDT

CHI 2018 Best Paper Award CHI 2018, April 21–26, 2018, Montréal, QC, Canada

Gender Recognition or Gender Reductionism? The Social Implications of Automatic Gender Recognition Systems

Foad Hamidi
University of Maryland,
Baltimore County (UMBC)
Baltimore, MD, USA
foadhamidi@umbc.edu

Morgan Klaus Scheuerman
University of Maryland,
Baltimore County (UMBC)
Baltimore, MD, USA
morgan.klaus@umbc.edu

Stacy M. Branham
University of Maryland,
Baltimore County (UMBC)
Baltimore, MD, USA
sbranham@umbc.edu

How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services

MORGAN KLAUS SCHEUERMAN, JACOB M. PAUL, and JED R. BRUBAKER, University of Colorado Boulder

The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition

OS KEYES, University of Washington, USA

ARGUMENT

Beijing's Big Brother Tech Needs African Faces

Zimbabwe is signing up for China's surveillance state, but its citizens will pay the price.

BY AMY HAWKINS

JULY 24, 2018, 10:39 AM

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.

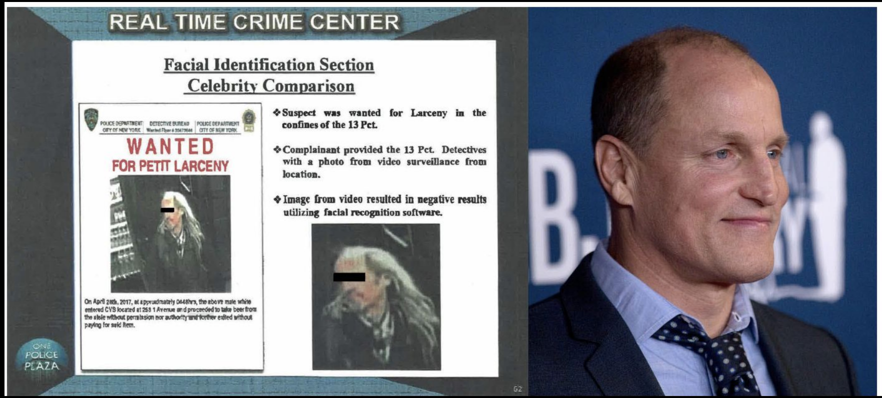
The Perpetual Line Up

(Garvie , Bedoya, Frankle 2016)

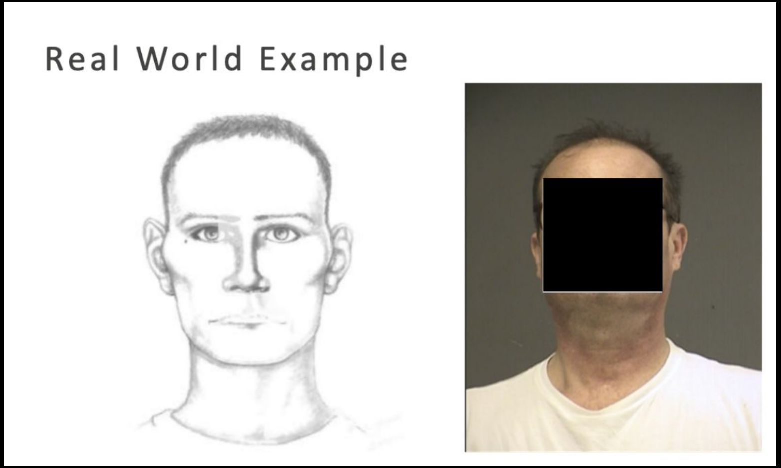
Facial Recognition is the Plutonium of AI

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

By Luke Stark



Celebrity faces as probe images



Composite sketches as probe images

[Garbage In, Garbage Out: Face Recognition on Flawed Data. Georgetown Law, Center on Privacy & Technology. www.flawedfacedata.com. 2019]

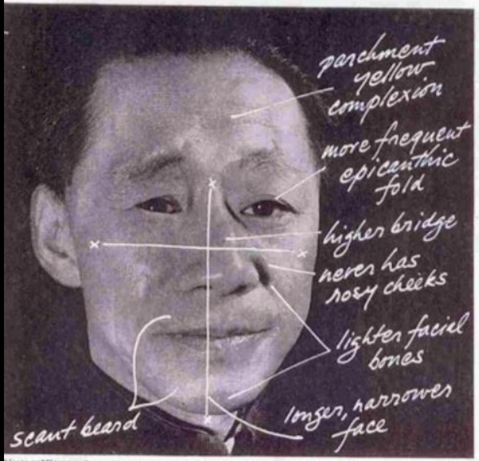
HOW TO TELL JAPS FROM THE CHINESE

ANGRY CITIZENS VICTIMIZE ALLIES WITH EMOTIONAL OUTBURST AT ENEMY

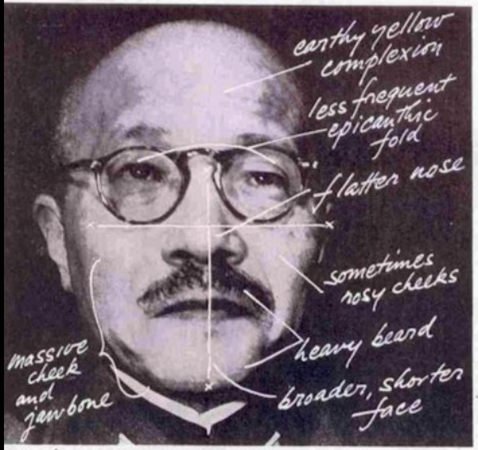
In the first discharge of emotions touched off by the Japanese assaults on their nation, U. S. citizens have been demonstrating a distressing ignorance on the delicate question of how to tell a Chinese from a Jap. Innocent victims in cities all over the country are many of the 75,000 U. S. Chinese, whose 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 anthroology, 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.



Chinese public servant, Ong Wen-hao, is representative of North Chinese anthropological group with long, fine-boned face and scant beard. Epicanthic fold of skin above eyelid is 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.



Japanese warrior, General Hideki Tojo, current Premier, is Samurai, closer to type of humble Jap than highbred relatives of Imperial Household. Typical are his heavy beard, massive 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.

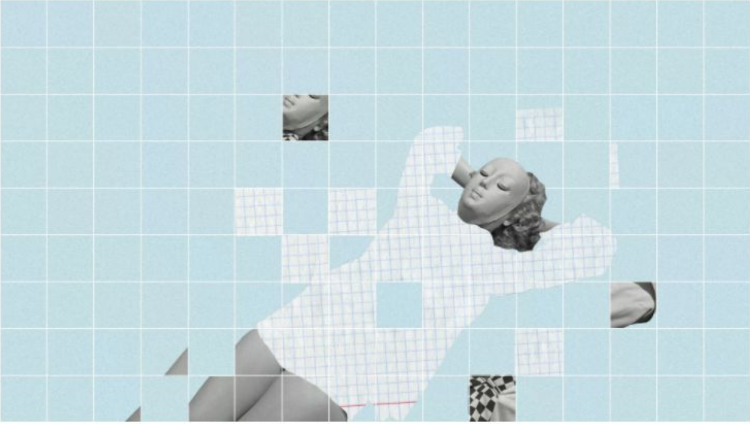
CONTINUED ON NEXT PAGE

One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority

In a major ethical leap for the tech world, Chinese start-ups have built algorithms that the government uses to track members of a largely Muslim minority group.

Fooling facial recognition with fashion

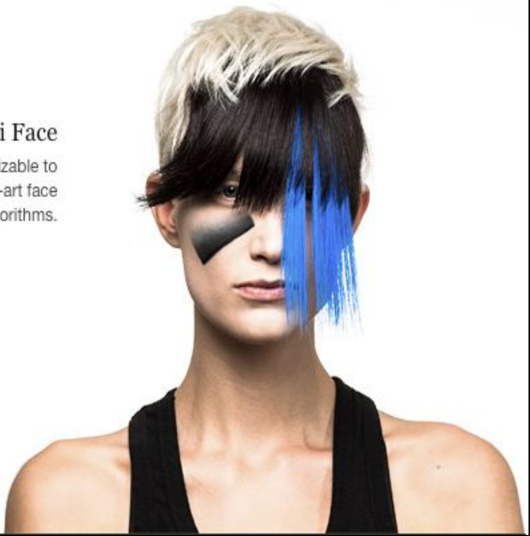
Jessie Li



by Coreana Museum of Art + Soobin Academy + G-square Model Academy

Anti Face

This face is unrecognizable to several state-of-art face detection algorithms.



Towards (more) socially responsible and ethics-informed research practices

Technology is not value-neutral

We are each accountable for the intended and unintended impacts of our work

Consider multiple direct and indirect stakeholders

Be attentive to the social relations and power differentials that shape construction and use of technology

I. Ethics-informed model testing

Comprehensive disaggregated evaluations:

- ❖ Compute metrics over subgroups defined along cultural, demographic, phenotypical lines
  - ➢ How you define groups will be context specific

- ❖ Consider multiple metrics - they each provide different information
  - ➢ Consider effects of different types of errors on different subgroups

		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

I. 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]

I. 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]

II. 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

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

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].

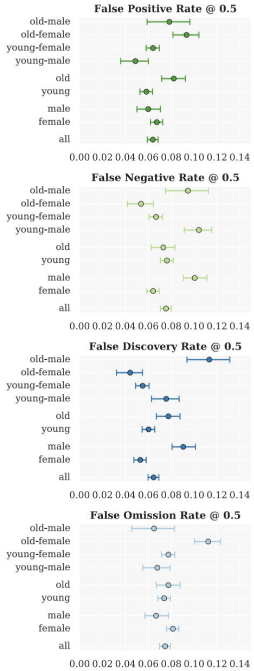
Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions 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.

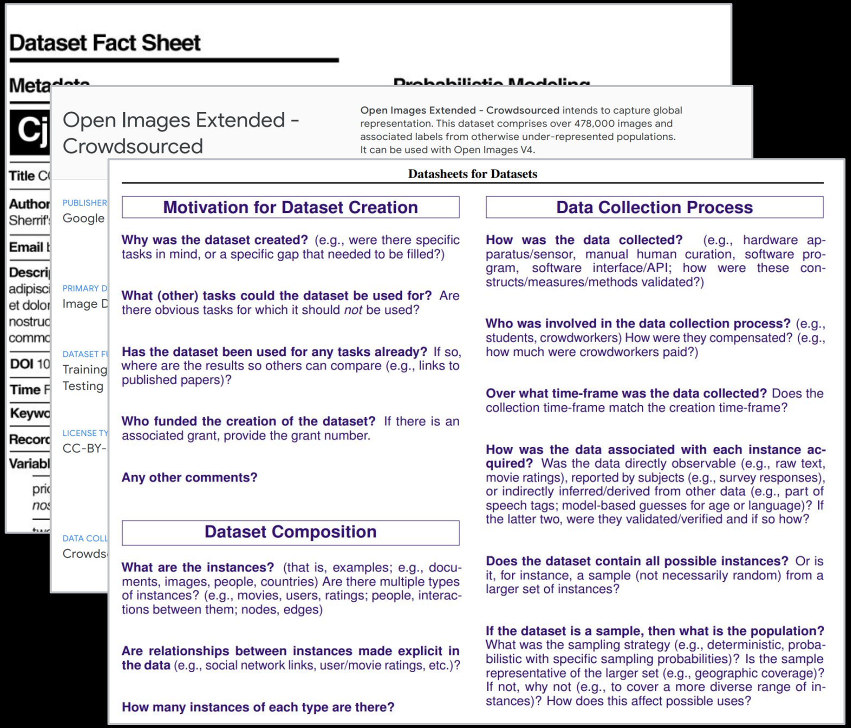
II. Model and data transparency

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]

III. Data is contingent, constructed, value-laden

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

III. 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

III. Data is contingent, constructed, value-laden

Shift how we think about data:

Data is fundamental to machine learning practice (not a means to an end)

Data should be considered a whole specialty in ML (Jo and Gebru, 2020)

Suggested readings:

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

Neff et al. (2017). [Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science]().

IV. Technology is not value-neutral

Technology is inherently political

As researchers and developers, we must shift our focus from intent ⟶ impact

~~“I’m just an engineer”~~

~~“I’m just doing basic research”~~

Suggested reading:

Green (2019). [Data Science as Political Action Grounding Data Science in a Politics of Justice]

Crawford et al. (2014). [Critiquing Big Data: Politics, Ethics, Epistemology]

V. Be attentive to your own positionality

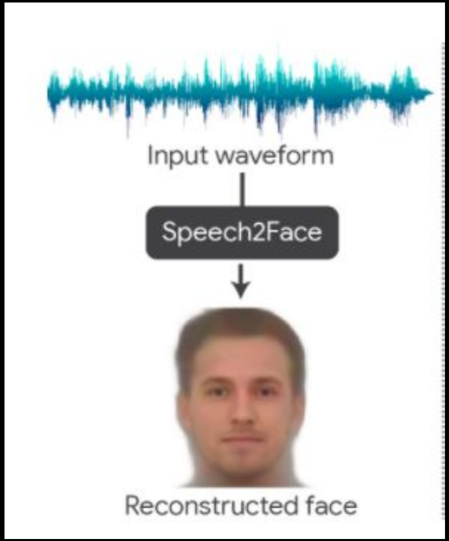
Our social positions in the world and set of experiences shapes and bounds our view of the world; this in turn affects the research questions we pursue and how we pursue them

Suggested readings:

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

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

V. Be attentive to your own positionality



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.

VI. Value knowledge and experience of marginalized groups

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

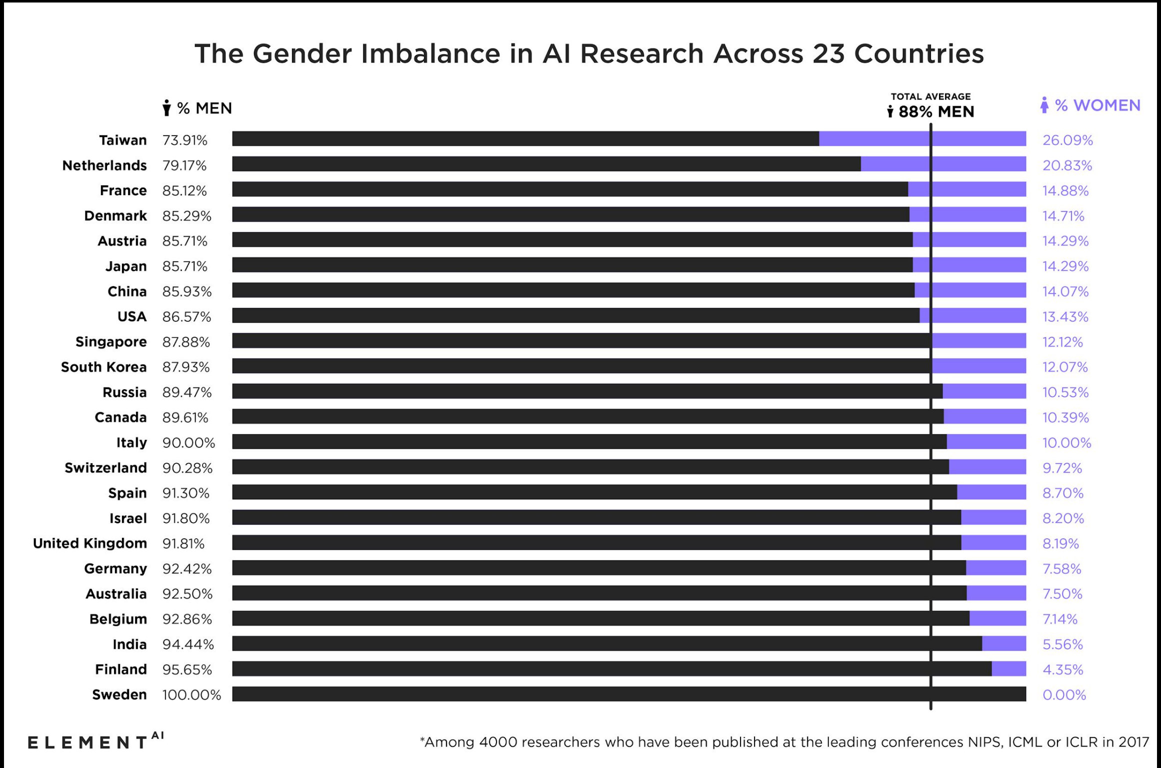
Sandra Harding (1991). Whose Science? Whose Knowledge?: Thinking from Women's Lives

VI. Value knowledge and experience of marginalized groups

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]



VI. Value knowledge and experience of marginalized groups

We can make intentional design choices to privilege the perspectives of marginalized stakeholders who are most at risk of being harmed by the technology we develop

Design Justice Network (www.designjustice.org)
Our Data Bodies (www.odbproject.org)

VII. Value interdisciplinarity and ‘non-technical’ work

Computer vision is simultaneously a technical and social discipline

Advancing racial literacy in tech

Different disciplinary practices give different types of knowledge

Non-technical work is valuable