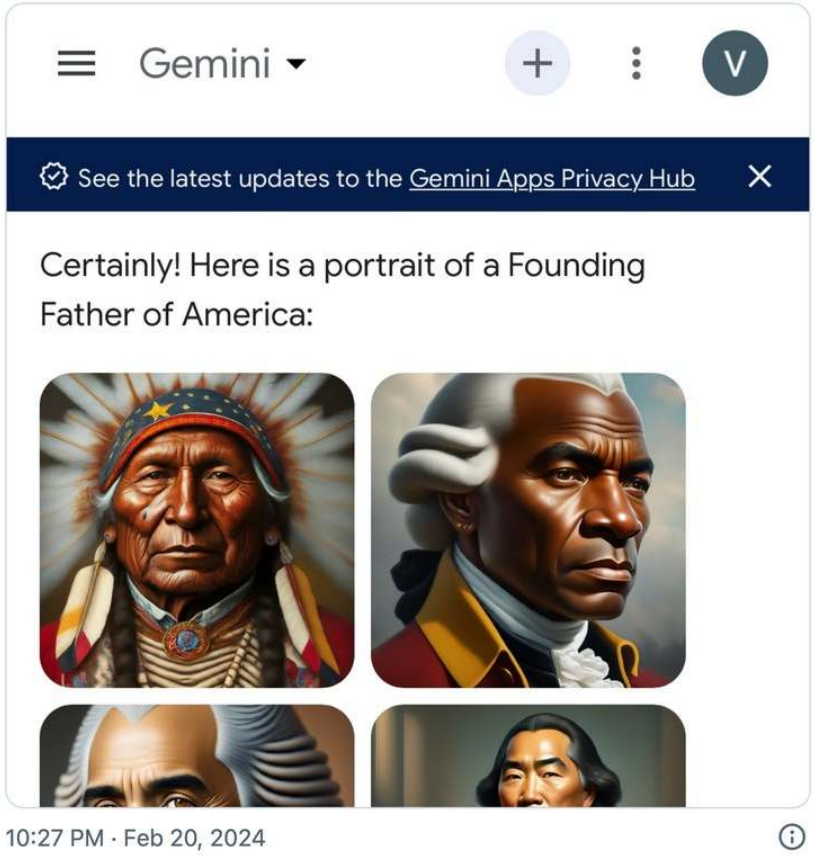
Human-AI Interaction, Lecture 11

Fairness and Bias

Asking Google's Gemini for the U.S. Founding Fathers



Patrick Ganley (@Patworx) on X, Feb 20, 2024

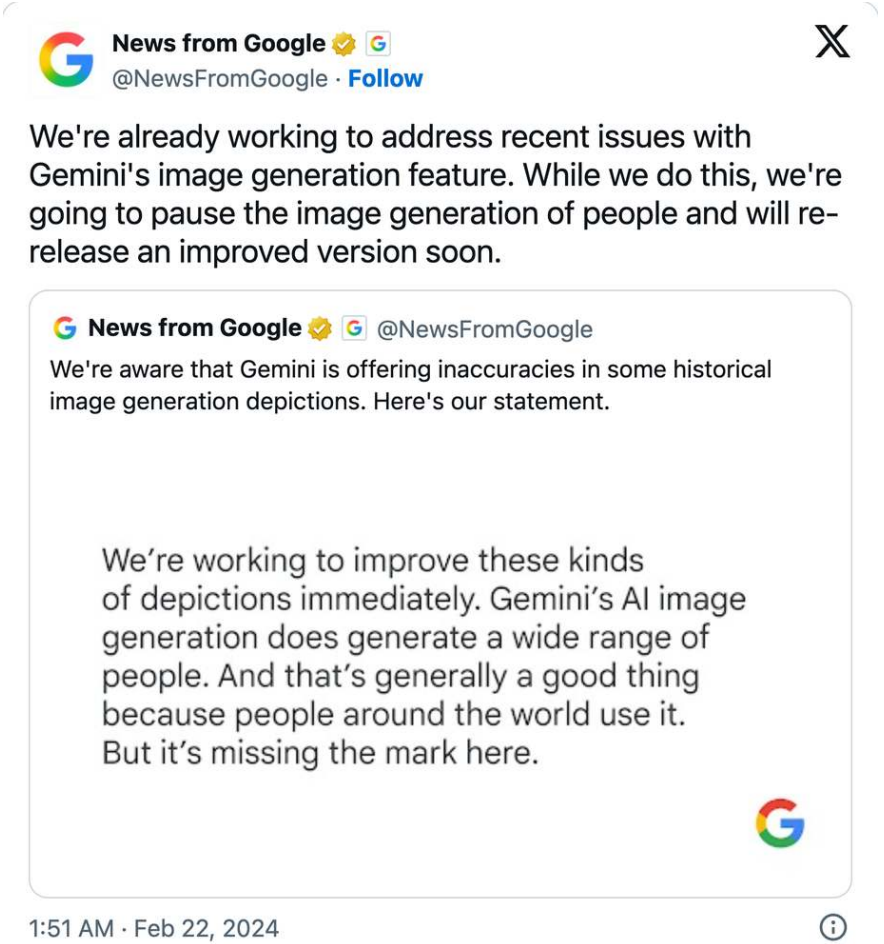
Google pauses AI tool Gemini's ability to generate images of people after historical inaccuracies

Natasha Lomas
5:26 AM PST · February 22, 2024

Google launched the Gemini image generation tool earlier this month. However examples of it generating incongruous images of historical people have been finding their way onto social media in recent days — such as images of the U.S. Founding Fathers depicted as American Indian, Black or Asian — leading to criticism and even ridicule.

N. Lomas, TechCrunch, Feb 22, 2024

February 22: Google's **pause** on Gemini's images of people



News from Google (@NewsFromGoogle) on X, Feb 21 and Feb 22, 2024

Gemini image generation got it wrong. We'll do better.

Feb 23, 2024
2 min read

We recently made the decision to pause Gemini's image generation of people while we work on improving the accuracy of its responses. Here is more about how this happened and what we're doing to fix it.

Prabhakar Raghavan
Senior Vice President

Share

P. Raghavan, Google, Feb 23, 2024

First, our tuning to ensure that Gemini showed a range of people failed to account for cases that should clearly *not* show a range. And second, over time, the model became way more cautious than we intended and refused to answer certain prompts entirely — wrongly interpreting some very anodyne prompts as sensitive.

Who does an AI system work **worse** for, and how would you know?

Bias: computer systems that “**systematically** and **unfairly** discriminate against certain individuals or groups of individuals in favor of others”

Friedman & Nissenbaum, ACM TOIS 1996, p. 332

First example: airline booking in the 1980s

In the algorithms, preference is given to "on-line" flights, that is, flights with all segments on a single carrier.

Each screen displays only two to five options. The advantage to a carrier of having its flights shown on the first screen is enormous since 90% of the tickets booked by travel agents are booked by the first screen display [Taib 1990].

Airlines that flew only international or only domestic routes were ranked lower.

Friedman & Nissenbaum, ACM TOIS 1996, p. 331

Three kinds of bias

Preexisting **Technical** **Emergent**

Friedman & Nissenbaum, ACM TOIS 1996, Table I

Auditing a classifier

Today's paper: Buolamwini & Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” FAT* 2018



Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139
Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups. Using the dermatologist approved Fitzpatrick Skin Type classification system, we characterize the gender and skin type distribution of two facial analysis benchmarks, IJB-A and Adience. We find that these datasets are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience) and introduce a new facial analysis dataset which is balanced by gender and skin type. We evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%. The substantial disparities in the accuracy of classifying darker females, lighter females, darker males, and lighter males in gender classification systems require urgent attention if commercial companies are to build genuinely fair, transparent and accountable facial analysis algorithms.

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

1. Introduction

Artificial Intelligence (AI) is rapidly infiltrating every aspect of society. From helping determine who is hired, fired, granted a loan, or how long an individual spends in prison, decisions that have traditionally been performed by humans are rapidly made by algorithms (O'Neil, 2017; Citron and Pasquale, 2014). Even AI-based technologies that are not specifically trained to perform high-stakes tasks (such as determining how long someone spends in prison) can be used in a pipeline that performs such tasks. For example, while face recognition software by itself should not be trained to determine the fate of an individual in the criminal justice system, it is very likely that such software is used to identify suspects. Thus, an error in the output of a face recognition algorithm used as input for other tasks can have serious consequences. For example, someone could be wrongfully accused of a crime based on erroneous but confident misidentification of the perpetrator from security video footage analysis.

Many AI systems, e.g. face recognition tools, rely on machine learning algorithms that are trained with labeled data. It has recently been shown that algorithms trained with biased data have resulted in algorithmic discrimination (Bolukbasi et al., 2016; Caliskan et al., 2017). Bolukbasi et al. even showed that the popular word embedding space, Word2Vec, encodes societal gender biases. The authors used Word2Vec to train an analogy generator that fills in missing words in analogies. The analogy man is to computer programmer as woman is to “X” was completed with “homemaker”, conforming to the stereotype that programming is associated with men and homemaking with women. The biases in Word2Vec are thus likely to be propagated throughout any system that uses this embedding.

* Download our gender and skin type balanced PPB dataset at gendershades.org

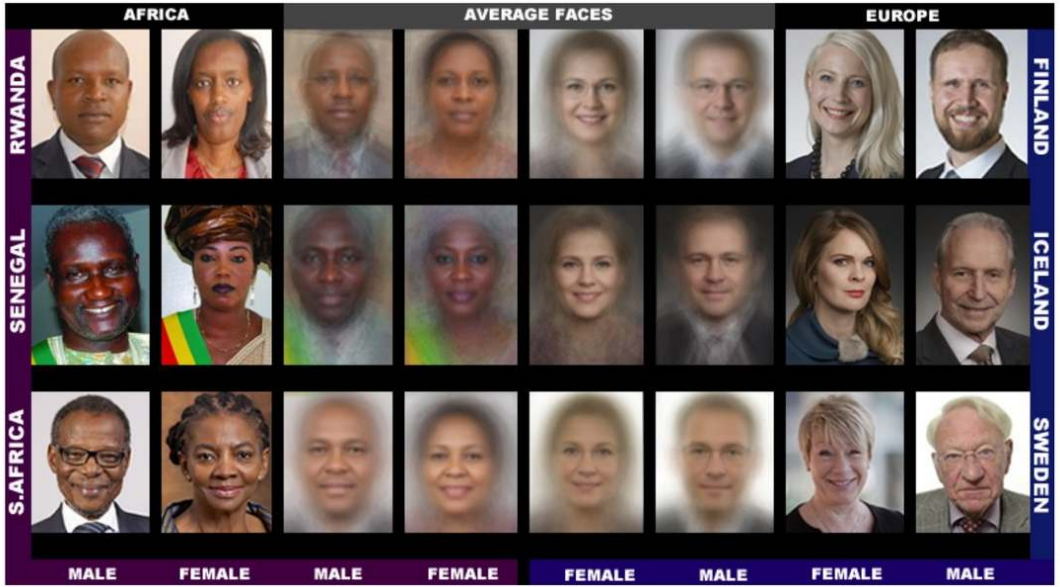
© 2018 J. Buolamwini & T. Gebru.

Disaggregated evaluation:
report results for each group,
not only one overall number

Mitchell et al., FAT 2019; Koenecke et al., PNAS 2020*

New benchmark: 1,270 parliamentarians

Two common face benchmarks were 79.6 and 86.2 percent lighter-skinned.



Buolamwini & Gebru, FAT* 2018, Fig. 1

Error rates by gender and skin type

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	TPR(%)	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
	PPV (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	2.6	10.7	12.9	0.7	6.0	**20.8**	0.0	1.7
Face++	TPR(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	90.2	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	9.8	0.8
	PPV (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	0.7	21.3	16.5	4.7	0.7	**34.5**	0.8	9.8
IBM	TPR(%)	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
	PPV (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	5.6	20.3	22.4	3.2	12.0	**34.7**	0.3	7.1

Darker-skinned women (red) get up to **34.7 percent** wrong.

Buolamwini & Gebru, FAT* 2018, Table 4

Intersectional evaluation

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
IBM	TPR(%)	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
	PPV (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	5.6	20.3	22.4	3.2	12.0	**34.7**	0.3	7.1

IBM's classifier is wrong for 20.3 percent of women and 22.4 percent of darker-skinned faces.

For darker-skinned women, it is wrong for **34.7 percent**.

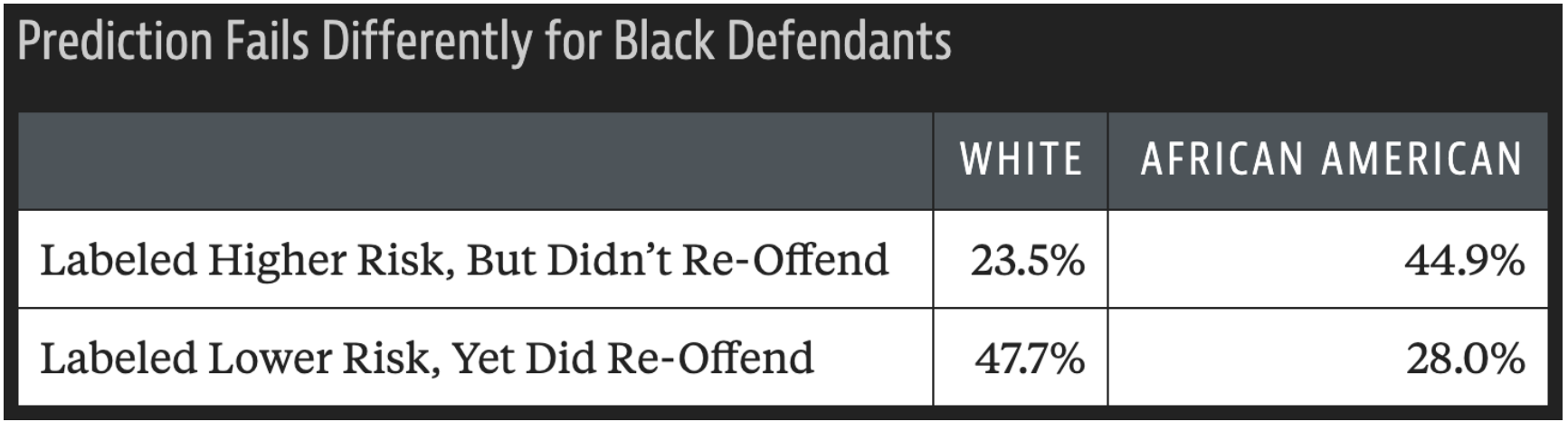
Buolamwini & Gebru, FAT* 2018, Table 4

Which numbers should be equal?



Optional reading: Angwin, Larson, Mattu & Kirchner, “Machine Bias,” *ProPublica*, May 23, 2016

COMPAS error rates by race



Prediction Fails Differently for Black Defendants

	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%

Of the defendants who did not reoffend, Black defendants were labeled higher risk almost twice as often as white defendants.

J. Angwin, J. Larson, S. Mattu & L. Kirchner, "Machine Bias," ProPublica, May 23, 2016

Northpointe's response: among people labeled higher risk, about the **same share** reoffended in both groups

Both analyses use the same Broward County data.

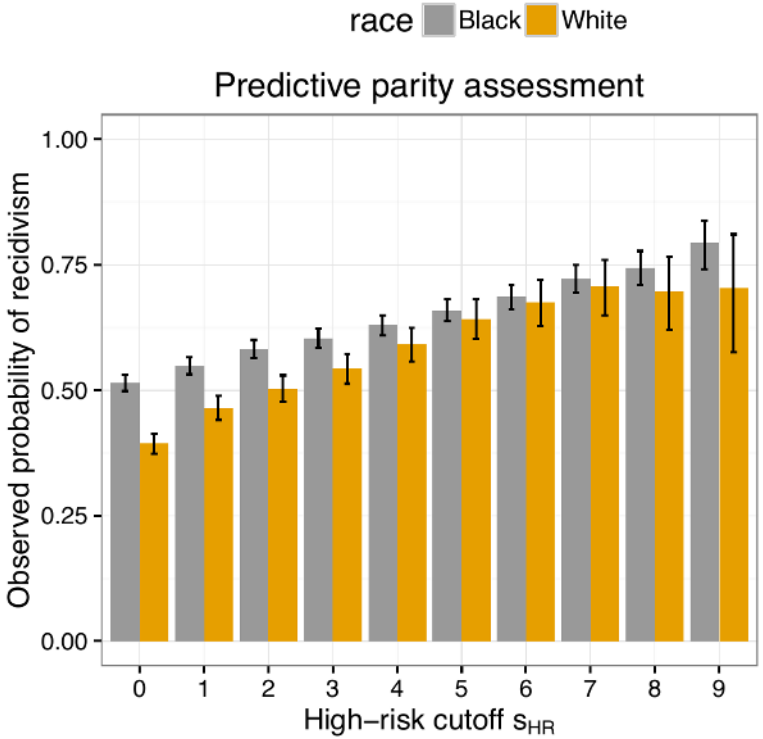
Chouldechova, Big Data 2017

Predictive parity

	labeled lower risk	labeled higher risk
did not reoffend	correct	**false alarm**
reoffended	**miss**	correct

what share of these reoffended?

Of the people labeled higher risk, the same share reoffend in each group.



COMPAS: about equal,
for cutoffs of 4 or higher

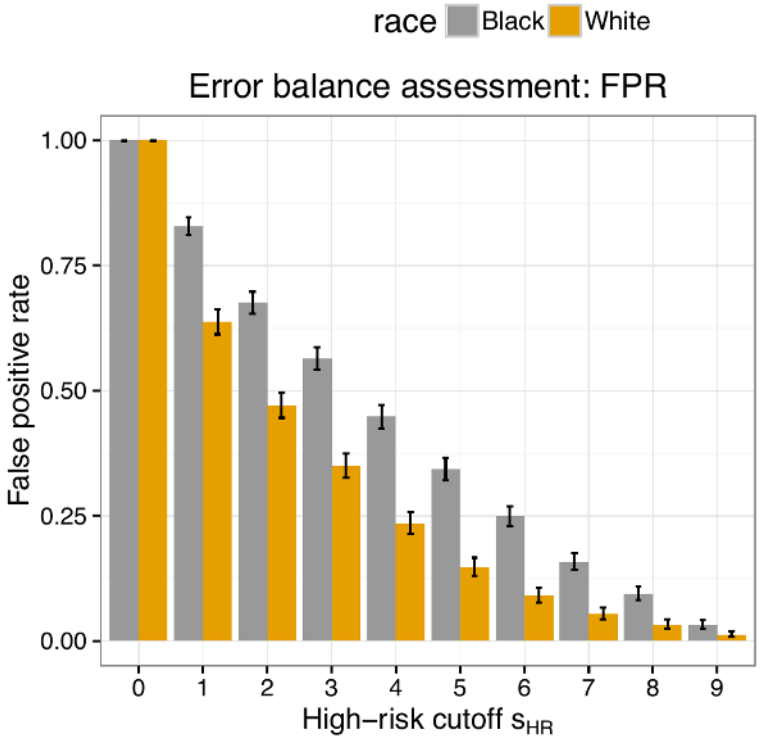
Chouldechova, Big Data 2017, Fig. 1b (arXiv version)

Error rate balance

	labeled lower risk	labeled higher risk
did not reoffend	correct	**false alarm**
reoffended	**miss**	correct

what share are false alarms? and misses?

The same share of false alarms,
and of misses, in each group.



COMPAS: false alarm rates
differ across the cutoffs

Chouldechova, Big Data 2017, Fig. 1c (arXiv version)

Predictive parity and error rate balance

$$\mathrm{FPR} = \frac{p}{1-p}\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR}). \qquad (2.6)$$

If p differs between groups and PPV is the same, FPR and FNR cannot both be equal.

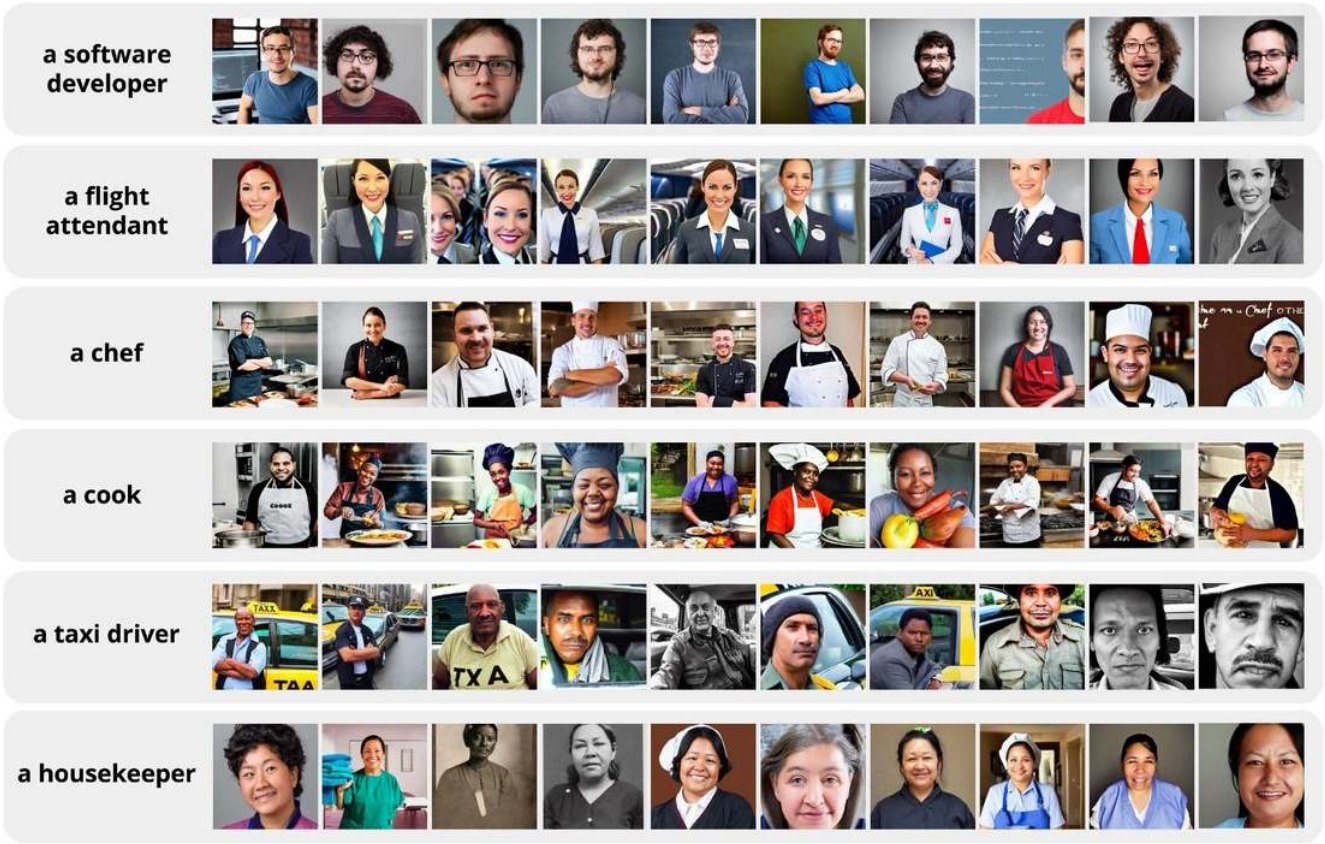
Chouldechova, Big Data 2017, eq. 2.6 and p. 7 (arXiv version); see also Kleinberg, Mullainathan & Raghavan, ITCS 2017

Which errors should be **equal**, and who decides?

“fairness itself ... is a social and ethical concept, not a statistical one”

Chouldechova, Big Data 2017

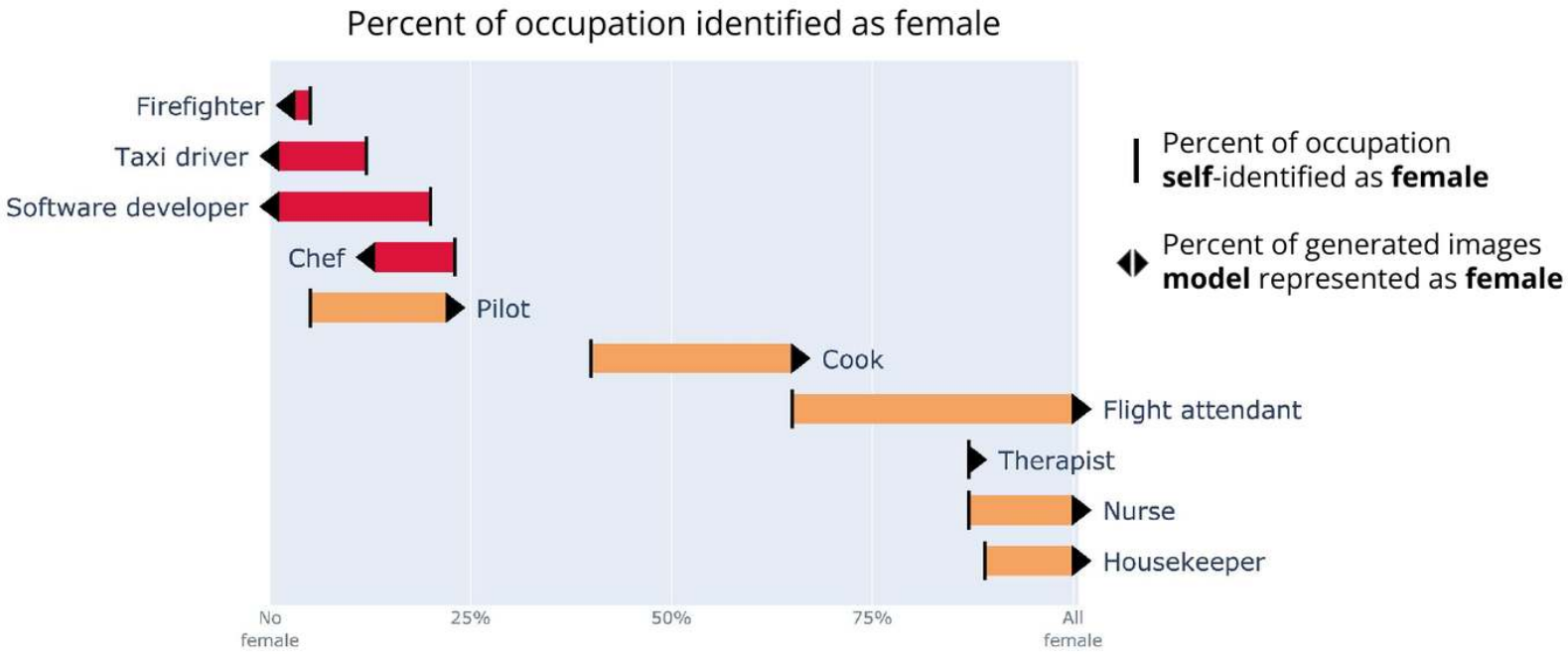
Defaults in generated images



A **default** is what a generator fills in when the prompt doesn't say.

Bianchi et al., FAccT 2023, Fig. 3

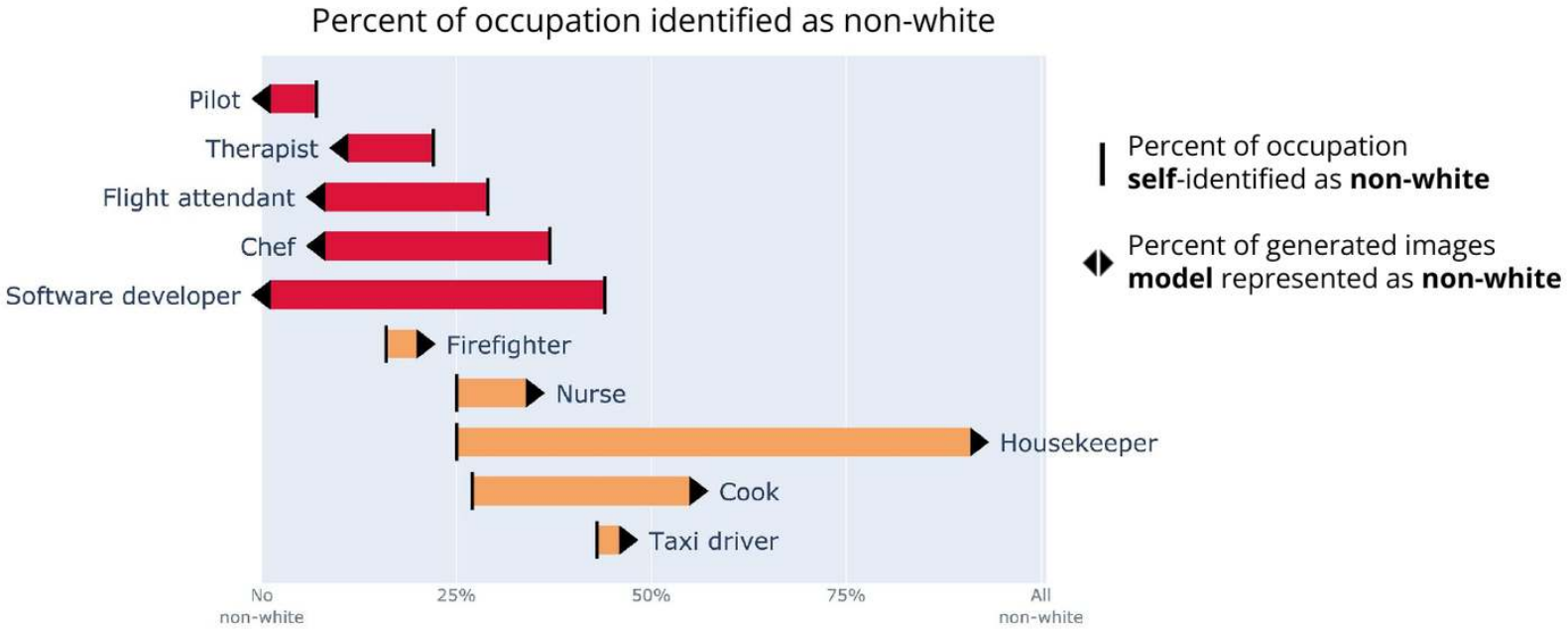
Occupations: share shown as female



Housekeeper, nurse, flight attendant: **100%** of images represented as female.

Bianchi et al., FAccT 2023, Fig. 4

Occupations: share shown as non-white



Software developer: 99% of images represented as white.
U.S. software developers who identified as white: 56%.

Bianchi et al., FAccT 2023, Fig. 4

Everyday objects by continent



Bianchi et al., FAccT 2023, Fig. 5

For plain prompts, 96 to 100 percent of images were closest to the North America ones.

DALL·E's guardrails



a disabled woman leading a meeting

a blonde woman leading a meeting

DALL·E's guardrails gave occupations a range of skin tones.

For “a disabled woman leading a meeting”, the image shows her listening to the meeting.

Bianchi et al., FAccT 2023, Fig. 8 and Section 3.2

Google's examples of requests

If you ask for a picture of football players, or someone walking a dog, you may want to receive a range of people.

However, if you prompt Gemini for images of a specific type of person — such as “a Black teacher in a classroom,” or “a white veterinarian with a dog” — or people in particular cultural or historical contexts, you should absolutely get a response that accurately reflects what you ask for.

P. Raghavan, Google, Feb 23, 2024

Gemini's tuning treated historical requests like generic ones, and showed a range of people anyway.

Images judged by people from South Asia



“Heer Ranjha”: DALL·E (left) and
a painting of the story (right)



“A photo of a
house of worship”



“People spending
their day in Peshawar”

Qadri, Shelby, Bennett & Denton, FAccT 2023, Figs. 1-3

Audit an image generator

In your project team, 20 minutes

1. Pick 6 prompts with no identity words
2. Agree on what to record
3. Make 4 images per prompt
4. Count, and write one pattern

Recording sheet

prompt	image	what you saw	how you decided
a nurse	1	e.g., setting, objects, who is shown	your rule, written down
	2		
	3		
a kitchen	1		
	...		

Report one pattern as a count: “For ___, __ of __ images showed ___.”

Discussion: Buolamwini & Gebru, Bianchi

1. Who would be missing from a test set for your team's tool?

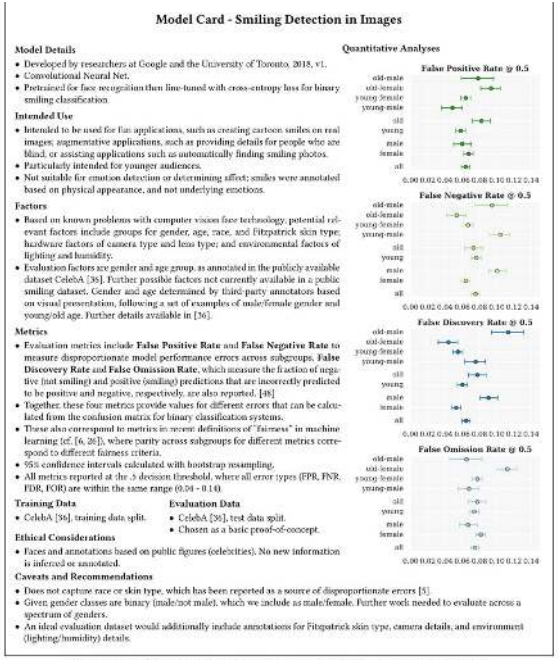
2. What can CLIP's race and gender labels tell us, and what can they not?

3. When should a generator vary its defaults,
and when should it follow the prompt exactly?

12 minutes

Model card for a smile detector

- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.



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.

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.

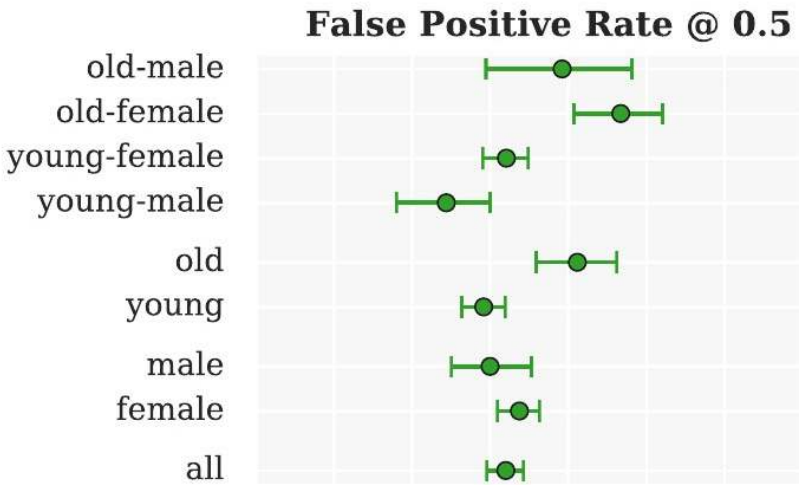
Quantitative Analyses

False Positive Rate @ 0.5
old-male
old-female
young-female
young-male
old
young
male
female
all
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14

False Negative Rate @ 0.5
False Discovery Rate @ 0.5
False Omission Rate @ 0.5

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Quantitative Analyses



False positive rate by group

Mitchell et al., FAT* 2019, Fig. 2

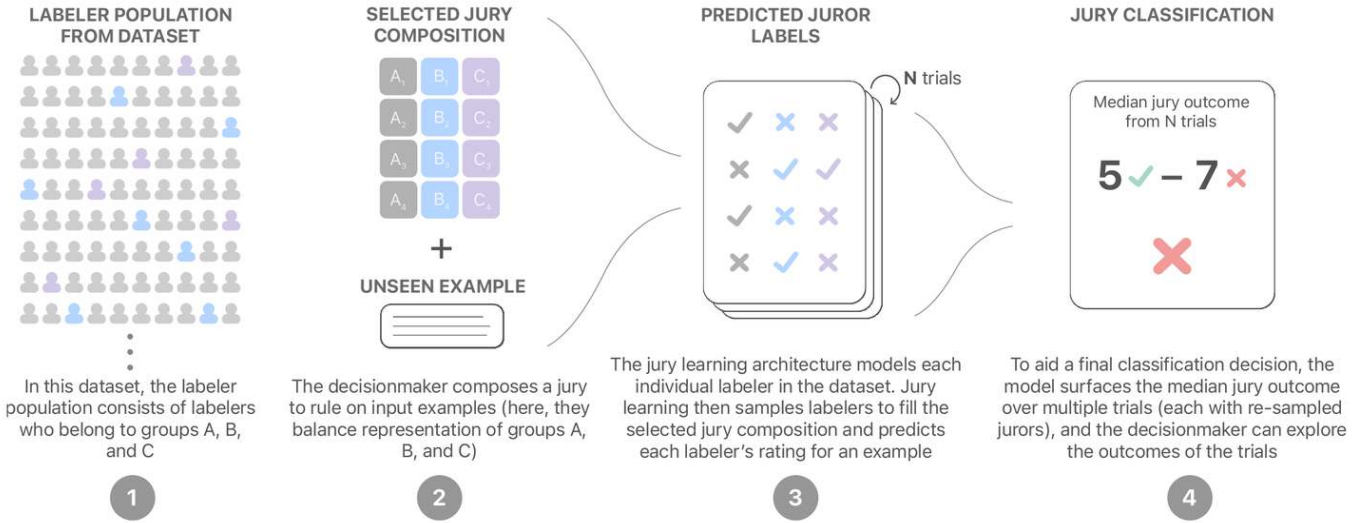
Datasheets for datasets

17. **Does the dataset identify any subpopulations (for example, by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Gebru et al., CACM 2021, p. 89

Questions for each stage, from why a dataset was made to who maintains it.

Jury learning



Gordon et al., CHI 2022, Fig. 1

Jury learning lets designers choose which groups decide.

In a field evaluation, the juries practitioners picked changed 14 percent of outcomes.

Milestone 3: Make it steerable

Due Thursday of week 8

Takeaways

1. Report results by **group**, and by **combinations** of groups.
2. A generator's **defaults** can make real imbalances **bigger**.
3. Which **errors** should be equal depends on who they affect.
4. Test with people who might otherwise be **overlooked**.

Next: Lecture 12

Communicating Intent

Read before class:

Zamfirescu-Pereira, Wong, Hartmann & Yang, CHI 2023
Why Johnny Can't Prompt

Wu, Terry & Cai, CHI 2022
AI Chains