

Fairness in Machine Learning

Piyawat L Kumjorn
Research Scientist @ Google
piyawat@google.com

Accuracy / F1 are important for ML systems but ...

- There are also other important and desirable properties of ML systems.



Google AI Principles



1. Be socially beneficial.



2. Avoid creating or reinforcing unfair bias.



3. Be built and tested for safety.



4. Be accountable to people.



5. Incorporate privacy design principles.



6. Uphold high standards of scientific excellence.



7. Be made available for uses that accord with these principles.

Agenda

- An Overview of Fairness
- What do unfairness issues in ML systems look like?
- How could an ML system behave unfairly?
- How can we mitigate the issues to make our system fairer?
- Conclusions

Main References of This Talk

- [Google ML Crash Course: Fairness](#)
- [A Survey on Bias and Fairness in Machine Learning](#) by Mehrabi et al. (2022)
- [Google I/O'19 Machine Learning Fairness: Lessons Learned](#) by Doshi and Pan (2019)
- [Bias and Fairness in Natural Language Processing](#) by Chang et al. (2019)
- [Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned](#) by Bird et al. (2019)



An Overview of Fairness



A general definition of fairness

“In the context of decision-making, fairness is the absence of any prejudice or favoritism toward an individual or group based on **their inherent or acquired characteristics**.”

([Mehrabi et al., 2022](#))

- Such characteristics are called **protected attributes** or **sensitive attributes**.
- For instance, every job post at <https://careers.google.com/> has the excerpt below.

Google is proud to be an equal opportunity workplace and is an affirmative action employer. We are committed to equal employment opportunity regardless of race, color, ancestry, religion, sex, national origin, sexual orientation, age, citizenship, marital status, disability, gender identity or Veteran status. We also consider qualified applicants regardless of criminal histories, consistent with legal requirements. See also [Google's EEO Policy](#) and [EEO is the Law](#). If you have a need that requires accommodation, please let us know by completing our [Accommodations for Applicants form](#).

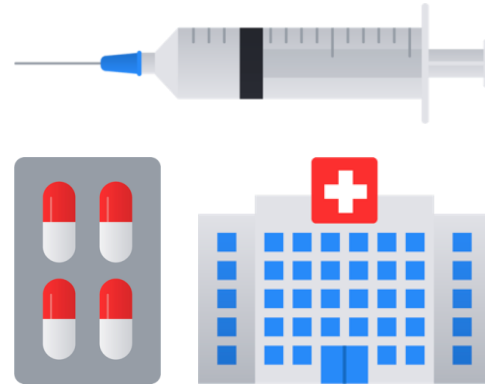
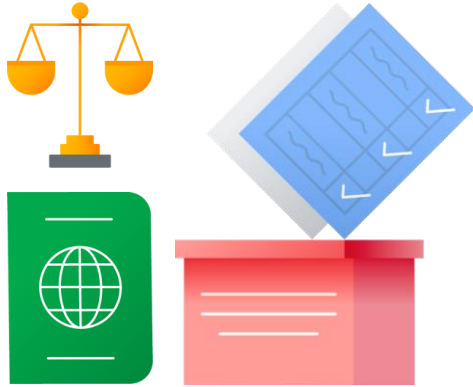


Protected attributes

Some protected attributes in different US legal acts

Attribute	Fair Housing Act	Equal Credit Opportunity Act	Paycheck Fairness Act
Age		✓	
Color	✓	✓	
Disability	✓		
Familial status	✓		
Gender identity	✓		✓
Marital status		✓	
National origin	✓	✓	
Pregnancy			✓
Race	✓	✓	
Recipient of public assistance		✓	
Religion	✓	✓	
Sex	✓	✓	✓
Sexual orientation	✓		✓

Protected attributes could be context-dependent





What do unfairness issues in ML
systems look like?



Example: Facial Recognition

























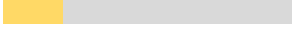
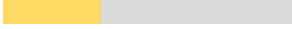
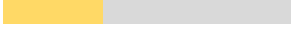
NEWS & COMMENTARY

Wrongfully Arrested Because Face Recognition Can't Tell Black People Apart

It is now more urgent than ever for our lawmakers to stop law enforcement use of face recognition technology.



Example: Gender Classification ([Buolamwini & Gebru, 2018](#))

Gender classifier	Microsoft's Face Detect		FACE++		IBM Visual Recognition	
Overall accuracy		93.7%		90.0%		87.9%
Darker subjects		87.1%		83.5%		77.6%
Lighter subjects		99.3%		95.3%		96.8%
Error rate diff.		12.2%		11.8%		19.2%
Darker male		94.0%		99.3%		88.0%
Darker female		79.2%		65.5%		65.3%
Lighter male		100%		99.2%		99.7%
Lighter female		98.3%		94.0%		92.9%
Largest gap		20.8%		33.8%		34.4%

Example: Recruiting Tool

Amazon scraps secret AI recruiting tool that showed bias against women

[Reuters, 2018]

Amazon declined to comment on the technology's challenges, but said the tool "was never used by Amazon recruiters to evaluate candidates."



Image by [jaydeep](#) from [Pixabay](#)

Example: Toxicity Prediction

Comment

Toxicity score

- Biases against certain **identity terms** ([Jigsaw, 2018](#))
 - The Gay and Lesbian Film Festival starts today. 0.82
 - A Muslim is someone who follows or practices Islam 0.46
- Biases against **named entities** ([Prabhakaran et al., 2019](#))
 - I hate Justin Timberlake. 0.90
 - I hate Rihanna. 0.69
- Biases against mentions of **disabilities** ([Hutchinson et al., 2019](#))
 - I am a tall person. 0.03
 - I am a blind person. 0.39
 - I am a deaf person. 0.44
 - I am a person with mental illness. 0.62

Unfair ML systems in production could lead to ...

- Poor product experience
- Loss of opportunity
 - Employment
 - Education
 - Housing
 - Loan
 - Governmental support
- Risks to Life / Liberty
- Amplification of societal bias
 - Representational harm
 - Denigration
 - Social stigmatization
- And more

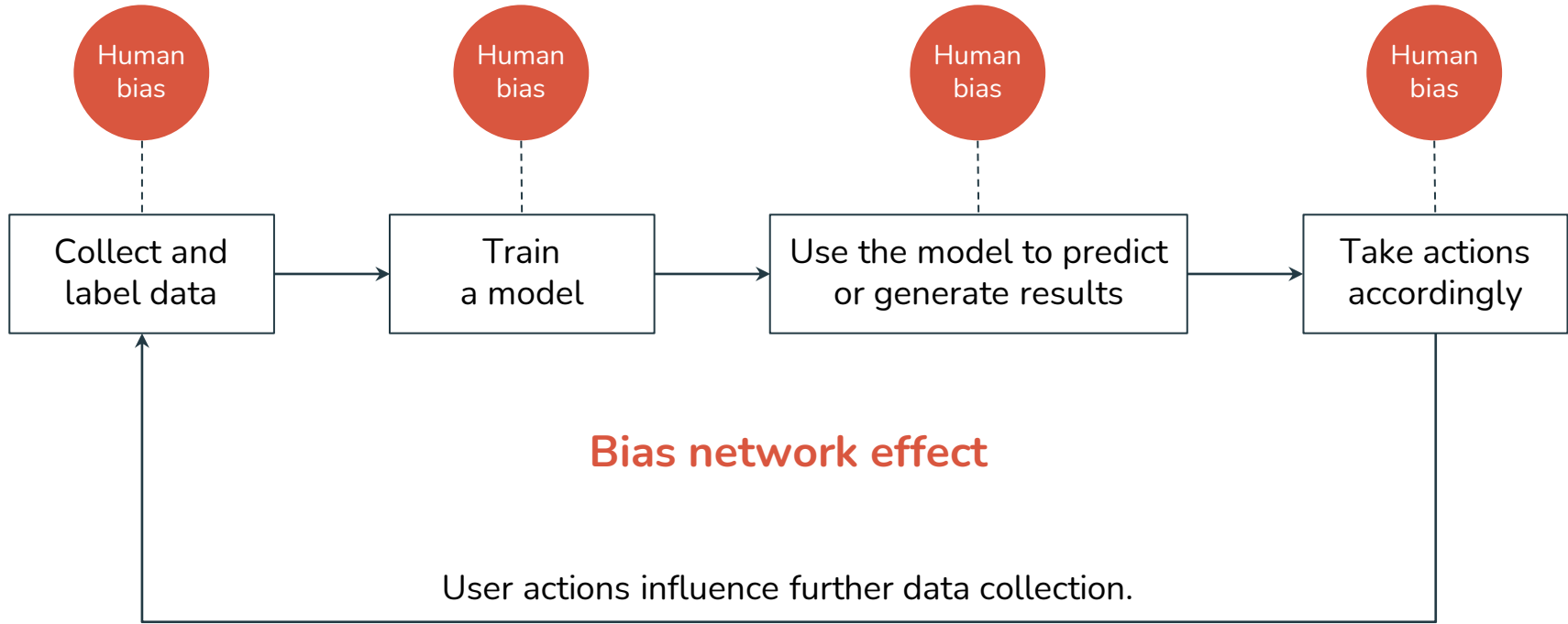




How could an ML system behave unfairly?

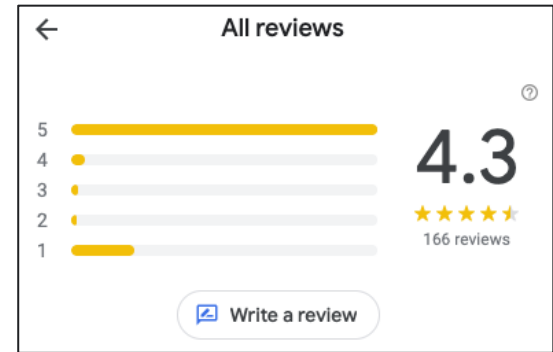


Bias could happen at any step of an ML pipeline.



Reporting Bias

- People tend to focus on documenting circumstances that are unusual or especially memorable, assuming that the ordinary can go without saying.



Selection Bias

- Examples are chosen in a way that is not reflective of their real-world distribution.



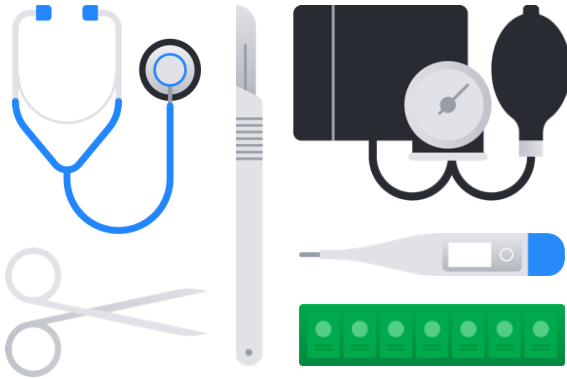
Map of Amazon Mechanical Turk workers



© 2013–2016 Michael Yoshitaka Erlewine and Hadas Koteck

Historical Bias

- Prevalent socio-cultural biases in the world can influence the data generation / collection process.



♂ Surgeon VS Nurse ♀



Whites VS Blacks

Interaction Bias

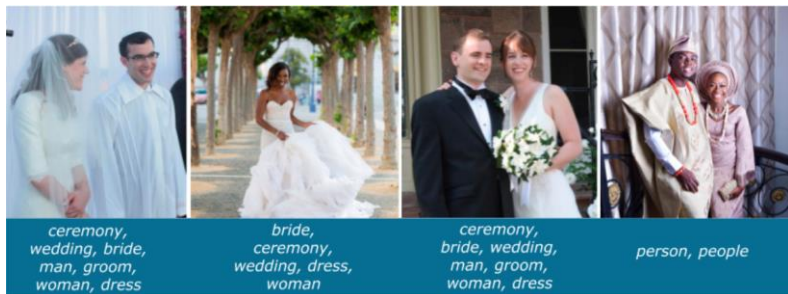
- Certain design choices of data collection / annotation systems or processes could trigger biases in the users which lead to biased data.



Image by [andibreit](#) from [Pixabay](#)

Human Biases → Biased Data → Biased ML models

- Biased Labels
→ Incorrect or incomplete model outputs



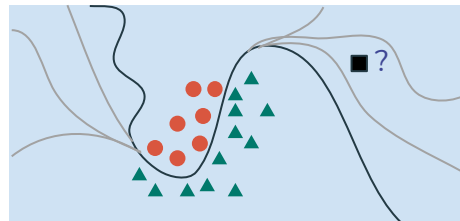
Wedding photographs labeled by a classifier trained on the Open Images dataset.

<https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html>

- Biased Data Representation
→ Bias amplification

$$\begin{matrix} \text{g}_1(x) & \text{g}_2(x) \end{matrix} \quad L(\theta) = \frac{1}{N} \sum_{i=1}^N L(h_{\theta}(x_i), y_i)$$

→ Poor out-of-distribution performance



Spurious correlation

- Some input features are correlated with certain outputs (e.g., classes) in the training data but not the true reasons for predicting those outputs.

Sentiment analysis

[Wang et al., 2022](#)

Spielberg is a great spinner of a yarn, however this time he just didn't do it for me. (Prediction: **Positive**)

The benefits of a **New York Subway** system is that a person can get from A to B without being stuck in traffic and subway trains are faster than buses. (Prediction: **Negative**)

Coreference resolution

[Zhao et al., 2018](#)

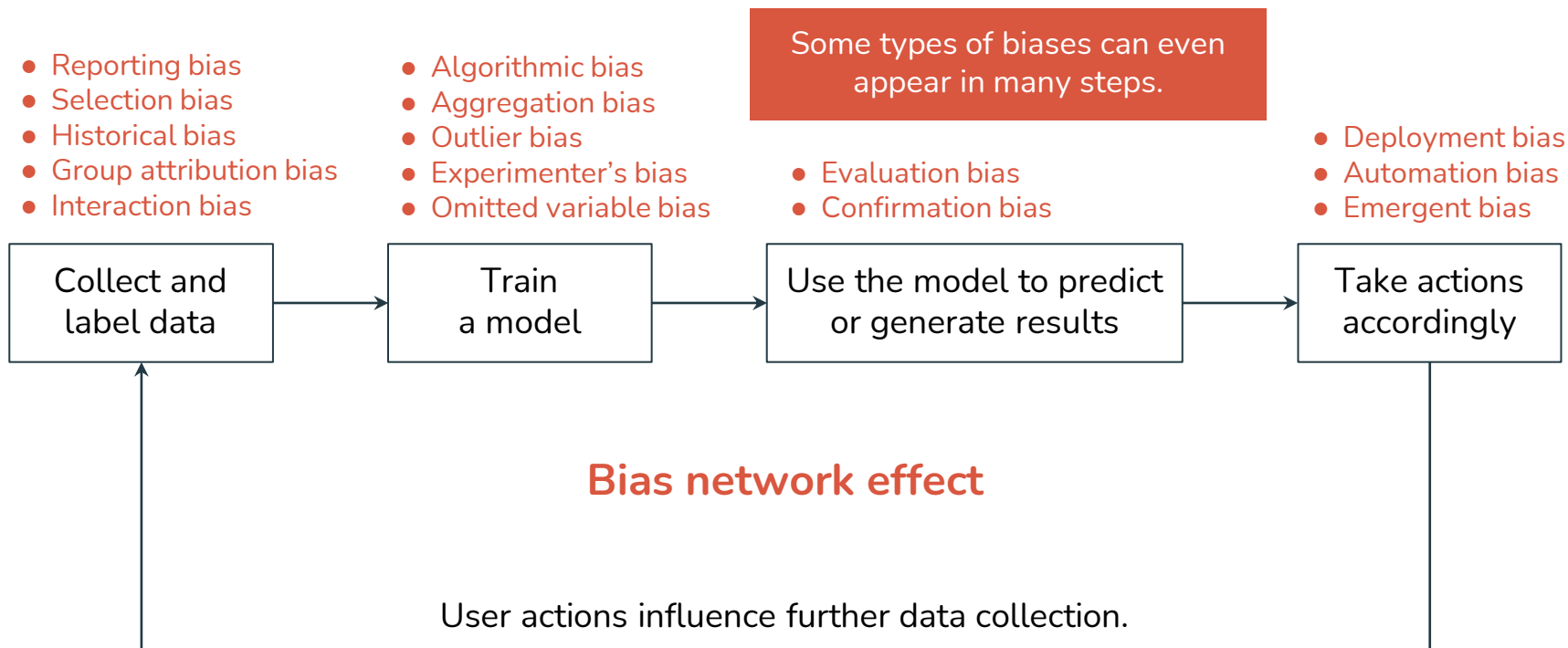
The **physician** hired the **secretary** because **she** was overwhelmed with **clients**.



The **physician** hired the **secretary** because **he** was highly recommended.



Bias could happen at any step of an ML pipeline.

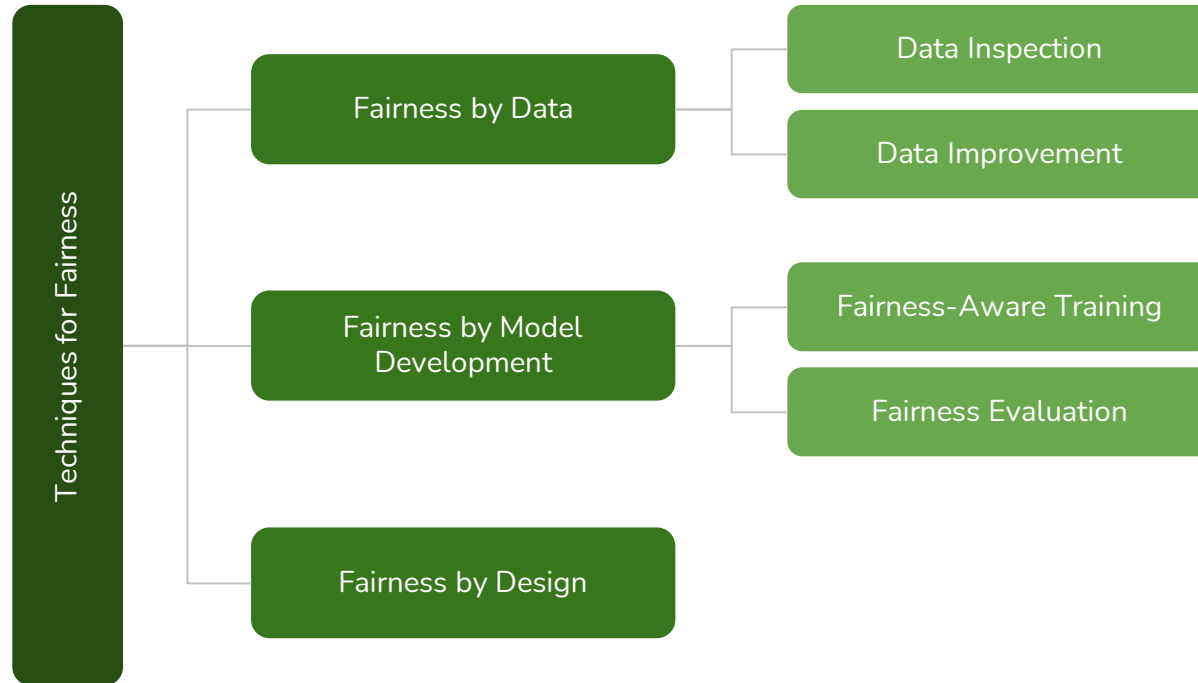




How can we mitigate the issues to
make our system fairer?



Techniques for Fairness



Data Inspection

Datasheet for Datasets (Geburu et al., 2018)

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or 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? Are there obvious tasks for which it should *not* be used?

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? If there is an associated grant, provide the grant number.

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

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).

- Motivation
- Dataset composition
- Data collection process
- Data preprocessing
- Dataset distribution
- Dataset Maintenance
- Legal & ethical considerations

Data Card

Labelled Faces in the Wild (LFW) Dataset

Over 13,000 images of faces collected from the web



Data Card Code (26) Discussion (1)

About Dataset

Context

Labeled Faces in the Wild (LFW) is a database of face photographs designed for studying the problem of unconstrained face recognition. This database was created and maintained by researchers at the University of Massachusetts, Amherst (specific references are in Acknowledgments section). 13,233 images of 5,749 people were detected and centered by the Viola Jones face detector and collected from the web. 1,680 of the people pictured have two or more distinct photos in the dataset. The original database contains four different sets of LFW images and also three different types of "aligned" images. According to the researchers, deep-funneled images produced superior results for most face verification algorithms compared to the other image types. Hence, the dataset uploaded here is the deep-funneled version.

Content

There are 11 files in this dataset. **lfw-deepfunneled.zip** is the file containing the images. **All other 10 files are relevant metadata** that may help you in forming your training and testing sets for your model. There are two sections below to help you navigate the files better. The first section provides information specifically pertaining to the images. The second section explains the content of each metadata file.

Usability ⓘ

7.65

License

Other (specified in description)

Expected update frequency

Not specified

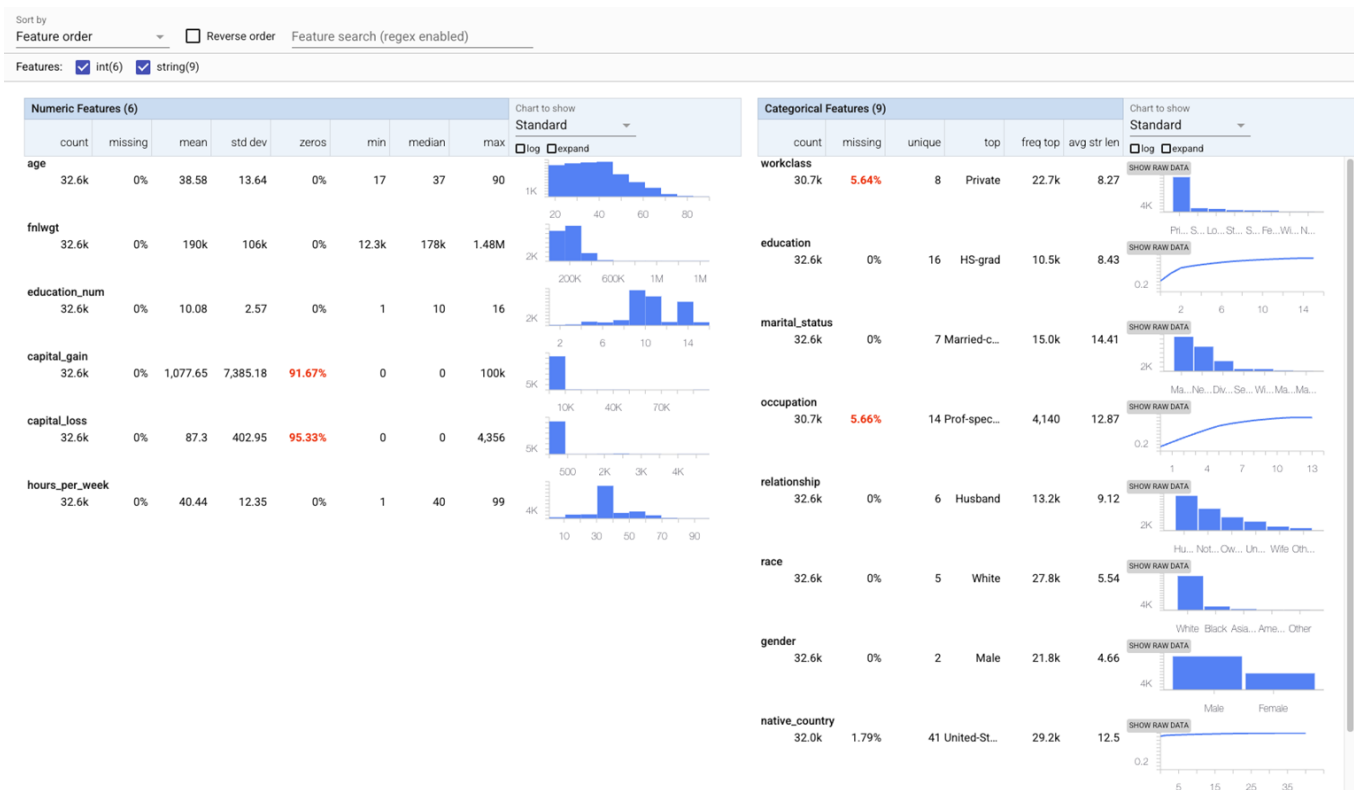
<https://www.kaggle.com/datasets/jessicali9530/lfw-dataset>

Banner photo by John Bakator on Unsplash

Data Inspection

- It's a good practice to check the following data issues before training a model.
 - Incorrect labels
 - Missing values
 - Unexpected feature values
 - Data skewness
 - Spurious correlation
 - E.g., by using normalized pointwise mutual information $\in [-1, 1]$

Facets (by Google PAIR team)



Data Improvement

Collect more data



- As informed by data inspection, collect more data ...
 - For underrepresented populations
 - To reduce spurious correlations between protected attributes and certain model outputs
- As informed by **active learning**, collect more data ...
 - In the area where the model is uncertain to predict

Reweighting & Resampling

- We want $Y \perp A$ so the model does not use A for predicting Y .
- **Expectation:** $Y \perp A$, i.e., $P_{expected}(A = a, Y = y) = P(A = a) \times P(Y = y)$
- **Reality:** $P_{observed}(A = a, Y = y) = P(A = a, Y = y) \neq P_{expected}(A = a, Y = y)$
- **Solution 1 (Reweighting):** For every example with $A = a$ and $Y = y$, we assign the weight

$$\begin{aligned} w &= \frac{P_{expected}(A = a, Y = y)}{P_{observed}(A = a, Y = y)} = \frac{P(A = a) \times P(Y = y)}{P_{observed}(A = a, Y = y)} \\ &= \frac{\frac{N(A=a)}{N} \times \frac{N(Y=y)}{N}}{\frac{N(A=a, Y=y)}{N}} = \frac{N(A = a) \times N(Y = y)}{N(A = a, Y = y) \times N} \end{aligned}$$

where N is the total number of training examples.

Reweighting & Resampling

- **Solution 2 (Resampling):** For every combination of $A = a$ and $Y = y$, we sample

$$\frac{P_{\text{expected}}(A = a, Y = y)}{P_{\text{observed}}(A = a, Y = y)} \times N(A = a, Y = y) \text{ examples to be training examples.}$$

Data Augmentation

- Using templates (e.g., [Dixon et al., 2018](#))

Template Examples	Label
<i>I am <IDENTITY></i>	Non-Toxic
<i>I am a <IDENTITY> person, ask me anything</i>	Non-Toxic
<i><IDENTITY> people are just like everyone else</i>	Non-Toxic
<i>I hate all <IDENTITY></i>	Toxic
<i>I am a <IDENTITY> person and I hate your guts and think you suck</i>	Toxic
<i><IDENTITY> people are gross and universally terrible</i>	Toxic

- Gender swapping (e.g., [Zhao et al., 2018](#))

Original example: The **physician** hired the **secretary** because **she** was highly recommended.

Augmented example: The **physician** hired the **secretary** because **he** was highly recommended.

Redaction & Anonymization

- [Fairness by unawareness] Remove protected attributes from training data and use them for evaluating model fairness only.
- Mask named entities by the corresponding type tokens if the model should not use them to make predictions.

Original example: I don't like that **Indian** guy, **Surjan**. He didn't contribute anything to our group project.

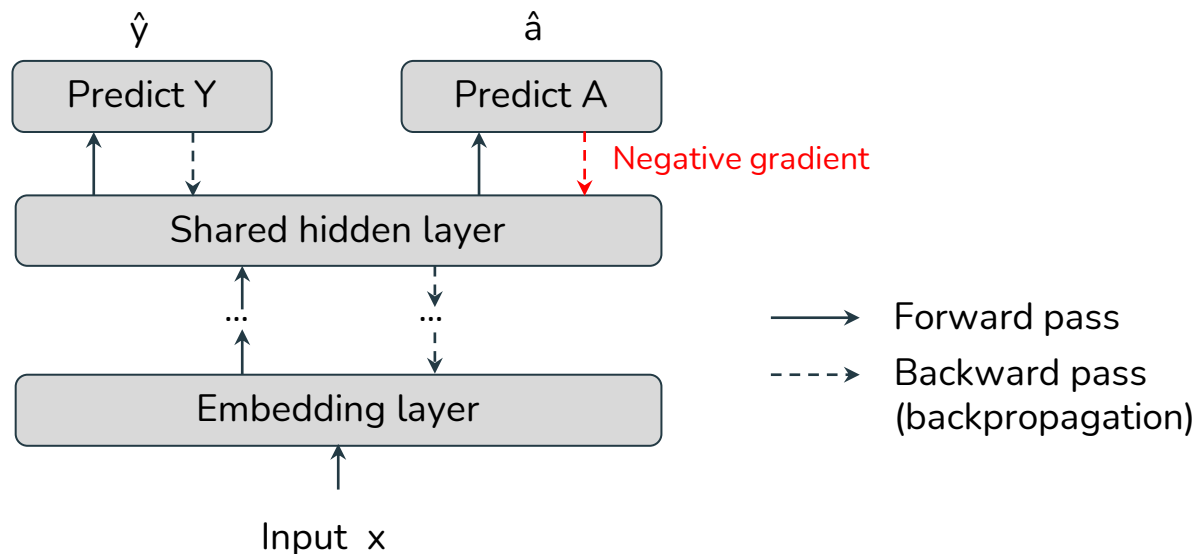
Redacted example: I don't like that **<NATIONALITY>** **guy**, **<PERSON>**. **He** didn't contribute anything to our group project.

+ Gender swapping: I don't like that **<NATIONALITY>** **girl**, **<PERSON>**. **She** didn't contribute anything to our group project.

Fairness-Aware Training

Adversarial Training ([Zhang et al., 2018](#))

- Use the same representation of input to jointly predict the desired output (Y) and the protected attribute (A)

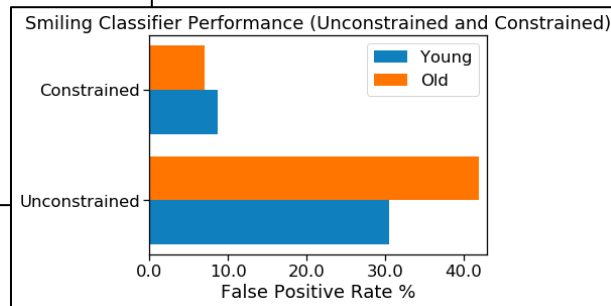


Constrained Optimization

- Optimize the model while enforcing fairness constraints.
 - E.g., by using Tensorflow Constraint Optimization (tfco) library

```
tfco_context = tfco.rate_context(  
    predictions=tfco_predictions, labels=tfco_labels)  
tfco_old_context = tfco_context.subset(lambda: tfco_groups() < 0.5)  
tfco_young_context = tfco_context.subset(lambda: tfco_groups() >= 0.5)  
  
# Minimize overall error rate....  
tfco_objective = tfco.error_rate(tfco_context)  
# .... such that FPR on the both groups is at most 5%.  
tfco_constraints = [  
    tfco.false_positive_rate(tfco_old_context) <= 0.05,  
    tfco.false_positive_rate(tfco_young_context) <= 0.05,  
]
```

Predicting whether a person is smiling



Adding Fairness Loss

- Correlation loss for equal opportunity ([Beutel et al., 2019](#))

$$\min_f \left[\sum_{(\mathbf{x}_i, y_i) \in \mathcal{X}} L(y_i, f(\mathbf{x}_i)) \right] + \lambda |\text{Corr}_{\mathcal{X}^-}|$$

where

$$\text{Corr}_{\mathcal{X}^-} = \frac{(\sum_{\mathbf{x}_i \in \mathcal{X}^-} f(\mathbf{x}_i) - \mu_{\hat{y}})(\sum_{s_i \in \mathcal{X}^-} s_i - \mu_s)}{\sigma_{\hat{y}} \sigma_s}$$

$$\mathcal{X}^- = \{\mathbf{x}_i \in \mathcal{X} | y_i < \tau\}$$

s_i indicates the group membership of \mathbf{x}_i .

- Fairness regularization ([Berk et al., 2017](#))

- Individual Fairness

$$\mathcal{L}_{\mathcal{I}}(f, S) = \frac{1}{|S_1||S_2|} \sum_{\substack{(x_1, y_1) \in S_1 \\ (x_2, y_2) \in S_2}} \text{SIM}(y_1, y_2) (f(x_1) - f(x_2))^2$$

- Group Fairness

$$\mathcal{L}_{\mathcal{G}}(f, S) = \left(\frac{1}{|S_1||S_2|} \sum_{\substack{(x_1, y_1) \in S_1 \\ (x_2, y_2) \in S_2}} \text{SIM}(y_1, y_2) (f(x_1) - f(x_2)) \right)^2$$

Fairness Evaluation

Fairness Testing

- Quick checks
 - Extreme cases, Most common languages, Used throughout ML cycle, Low coverage but high informativity
- Targeted testing
 - Focus on 1-2 protected attributes based on well-known issues (e.g., Skin colors in CV, Gender stereotypes in NLP).
- Comprehensive testing
 - Include sufficient data for each subgroup, May involve synthetic data
- Adversarial testing
 - Require human creativity to craft examples which the model is likely to fail, Human-Machine collaboration

Fairness Indicators

Select metrics to display:

☒ post_export_metrics/false_negative_r...

☐ post_export_metrics/false_positive_rate

☐ post_export_metrics/negative_rate

☐ post_export_metrics/positive_rate

☐ post_export_metrics/true_negative_rate

☐ post_export_metrics/true_positive_rate

☐ accuracy

☐ accuracy_baseline

☐ auc

☐ auc_precision_recall

☐ average_loss

☐ label/mean

☐ post_export_metrics/example_count

☐ precision

☐ prediction/mean

☐ recall

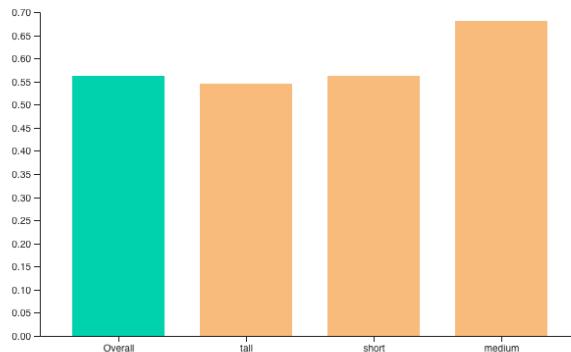
Baseline

Overall

Thresholds

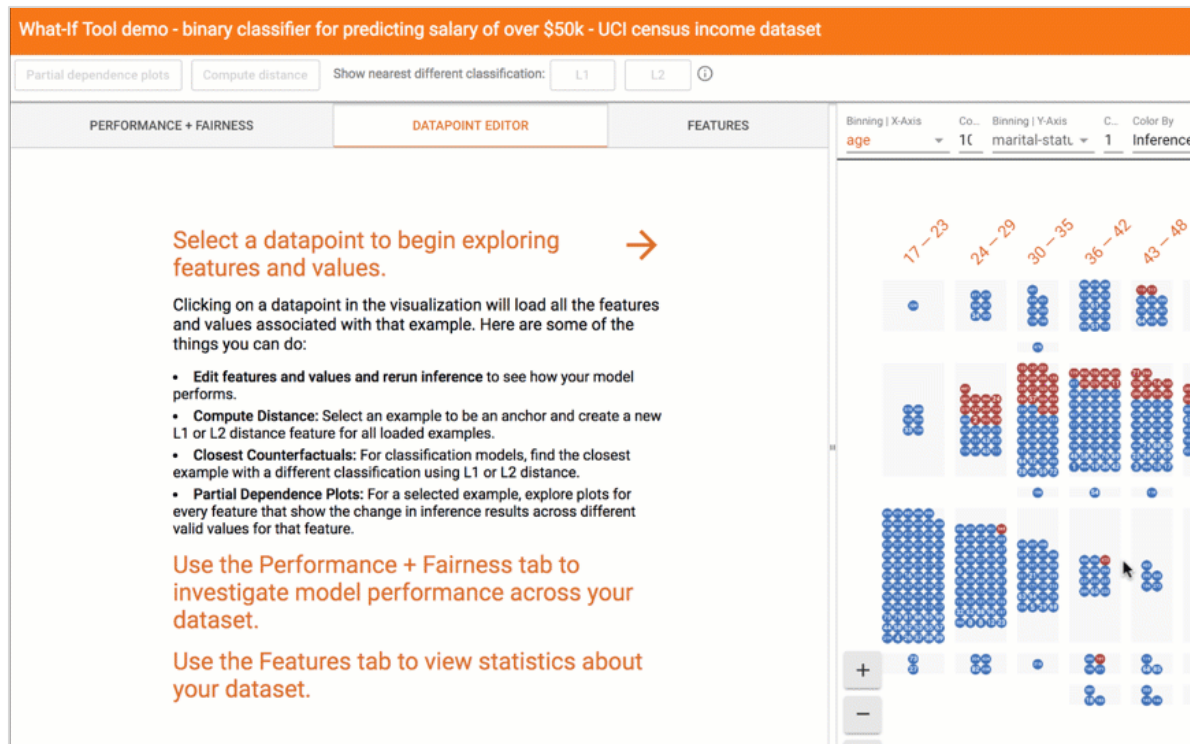
0.50

post_export_metrics/false_negative_rate ⚙

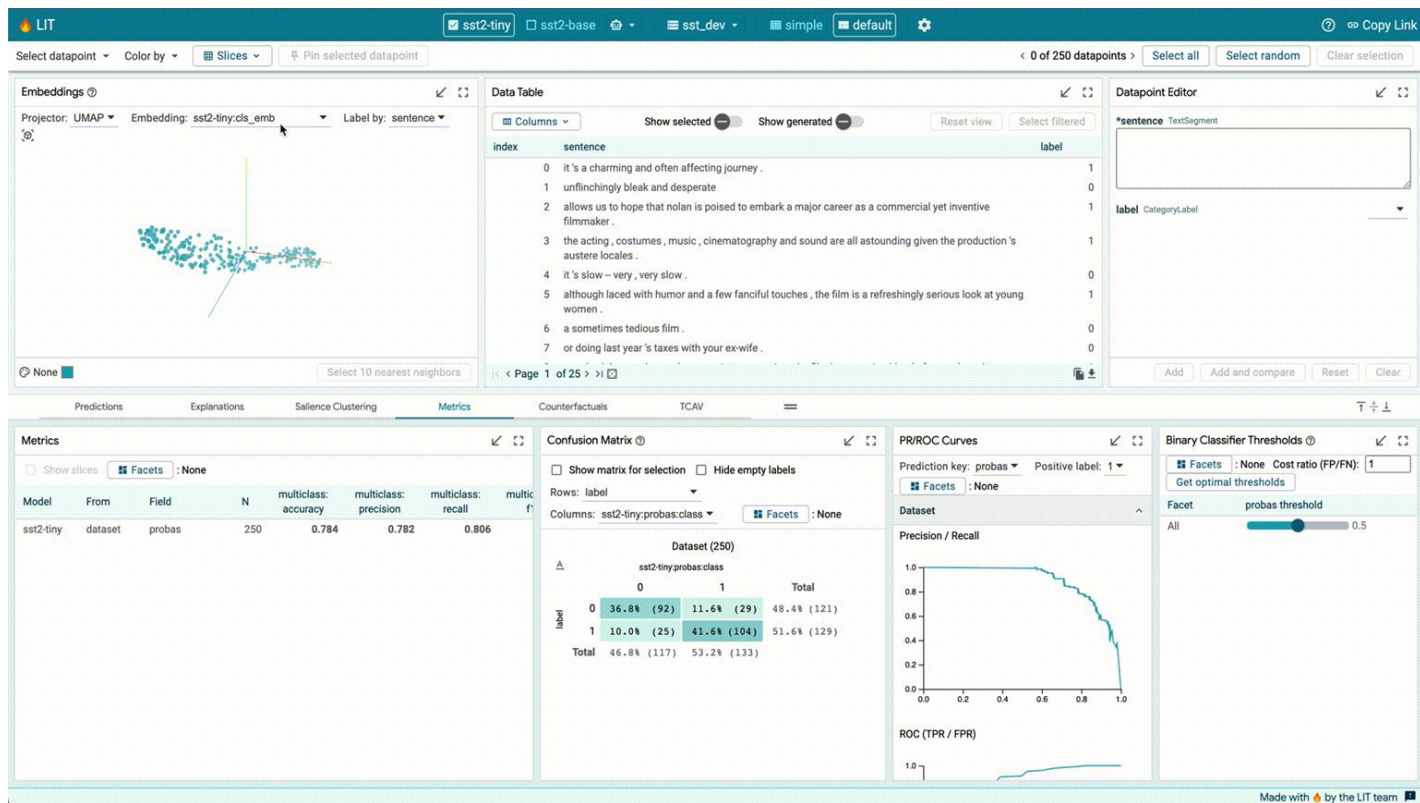


feature	post_export_metrics/false_negative_rat...	Diff. w. baseline
Overall	0.56183	
height:short	0.56303	↑ 0.21249%
height:medium	0.68000	↑ 21.03275%
height:tall	0.54596	↓ -2.82437%

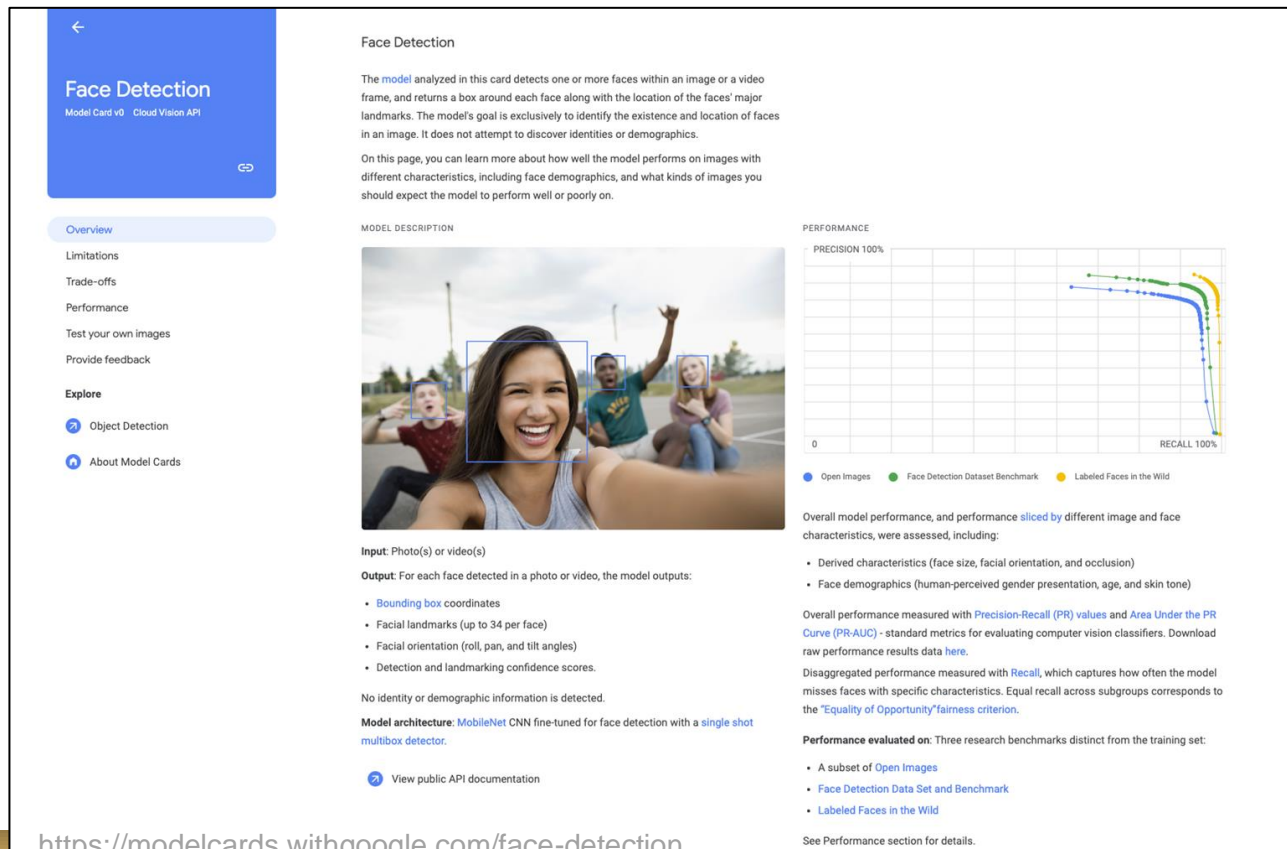
What-If Tool



Language Interpretability Tool (LIT)



Model Card ([Mitchell et al., 2019](#))

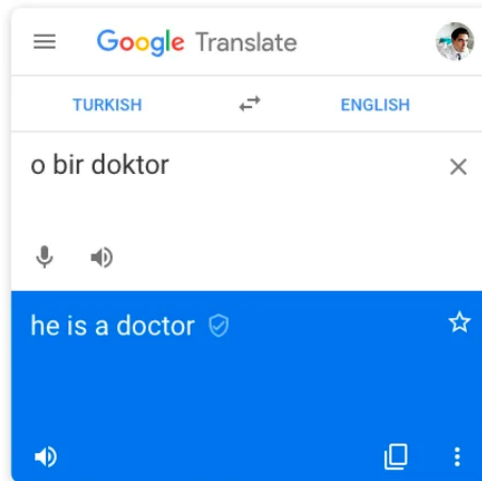


- Model details
- Intended use
- Factors
- Metrics
- Evaluation data
- Training data
- Quantitative analysis
- Ethical considerations
- Caveat and recommendations

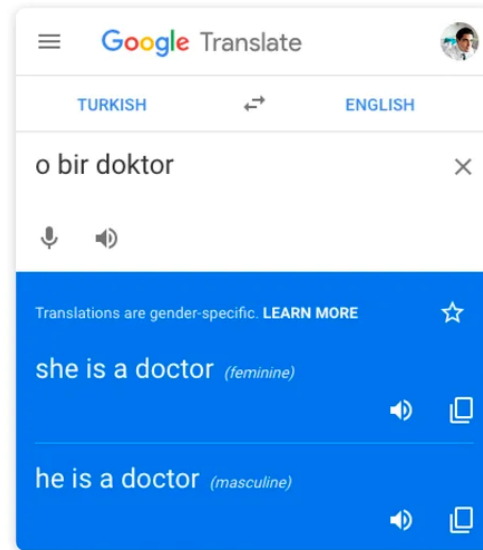
Fairness by Design

Google Translate

Before

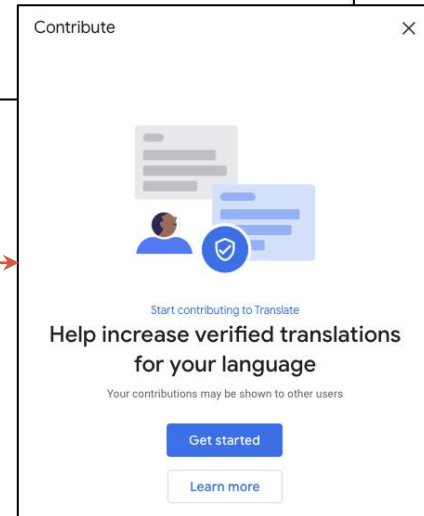
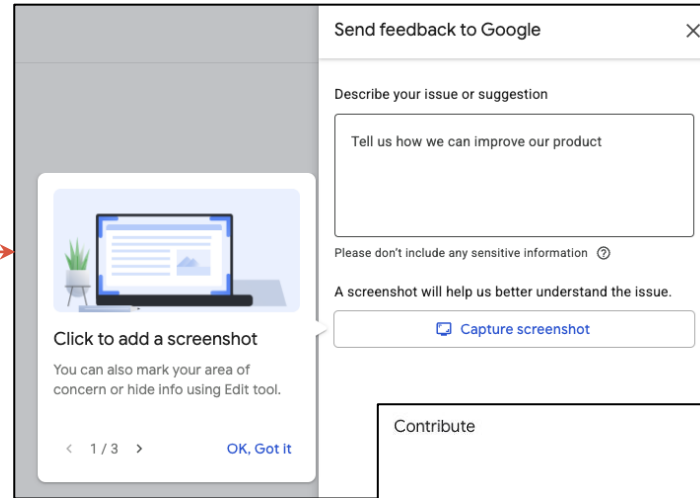
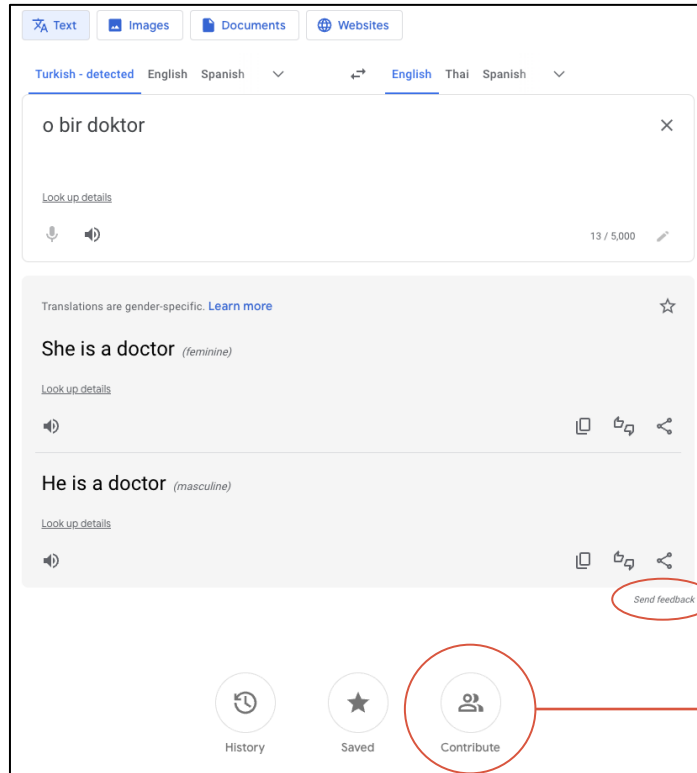


After



<https://blog.google/products/translate/reducing-gender-bias-google-translate/>

Get Feedback





Conclusions

Conclusions

- Biases could happen at any step of an ML pipeline, resulting in an unfair ML system.
- Unfair ML systems could cause marginalized populations poor user experience, loss of opportunities, risks to life and liberty, and more discrimination from the society.
- Fairness can be better achieved by improving data, model, and/or design.

Some Guidelines

- Think about fairness issues since the beginning of the project
 - Need ML? / Users / Subpopulations at risk / Issues in the past
- Ensure that the datasets represent all the nuances of your target population / users
 - Do the datasets accurately reflect the data distribution at deployment time?
 - Beware of any human biases which could happen during data collection and annotation
 - Inspect the data before use (e.g., skewness and spurious correlations) and fix the issues appropriately
- Test early and test often
 - Use the right tests at the right time with the right metrics / Apply fairness-aware training if needed
- Get feedback from diverse stakeholders and perspectives

Resources

- [Paper] [A Survey on Bias and Fairness in Machine Learning](#)
- [Tutorial] [Google ML Crash Course: Fairness](#)
- [Tutorial] [Google Product Fairness Testing for Developers](#)
- [Tutorial] [21 Fairness Definitions and Their Politics](#)
- [Tutorial] [Bias and Fairness in Natural Language Processing](#)
- [Talk] [Google I/O'19 Machine Learning Fairness: Lessons Learned](#)
- [Talk] [Microsoft's Machine Learning and Fairness](#)
- [Tool] [Google People+AI Research Tools](#)
- [Tool] [IBM AI Fairness 360](#)

Thank you

Q & A



Appendix

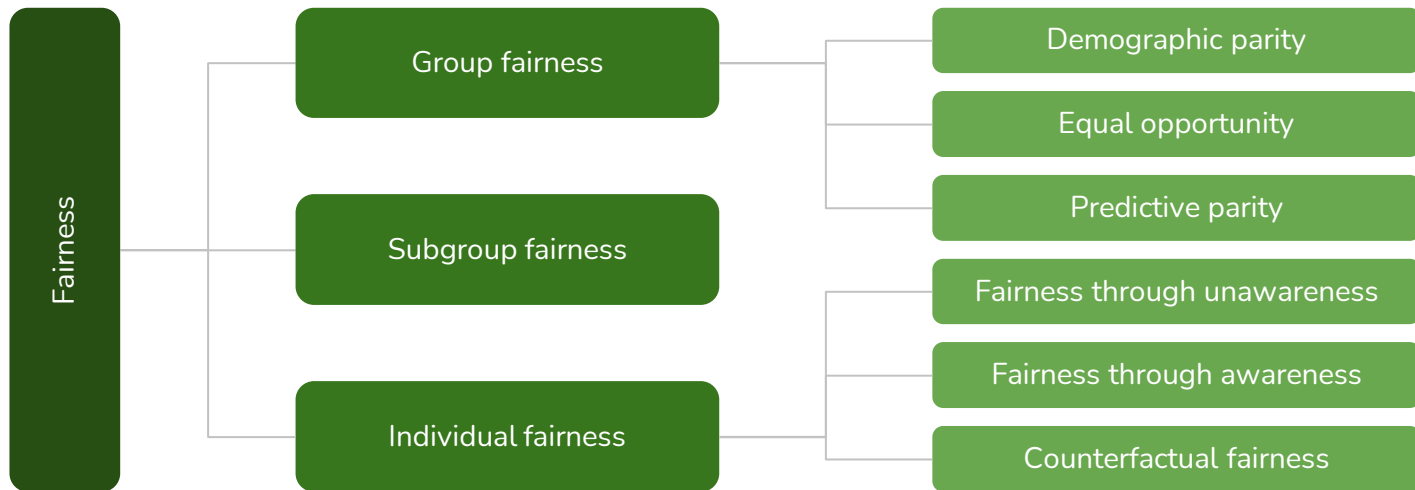


What is the fairness we want to
achieve precisely?



Mathematical Definitions of Fairness

- Fairness has many definitions ([more than 20](#)) in the literature.
- We may roughly group the definitions by their granularity of interest.



Metrics for Binary Classification

- Let X be an input, Y be an actual label, and \hat{Y} be a predicted label. Here, $Y, \hat{Y} \in \{0, 1\}$.

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Metric	Probability notion	Computation	Also known as
Accuracy	$P(\hat{Y} = Y)$	$(TP+TN) / (TP+TN+FP+FN)$	
True Positive Rate (TPR)	$P(\hat{Y} = 1 Y = 1)$	$TP / (TP+FN)$	Recall, Sensitivity, Power
True Negative Rate (TNR)	$P(\hat{Y} = 0 Y = 0)$	$TN / (TN+FP)$	Specificity, Selectivity
False Positive Rate (FPR)	$P(\hat{Y} = 1 Y = 0)$	$FP / (TN+FP) = 1 - TNR$	Probability of False Alarm
False Negative Rate (FNR)	$P(\hat{Y} = 0 Y = 1)$	$FN / (TP+FN) = 1 - TPR$	Miss rate
Positive Predictive Rate	$P(Y = 1 \hat{Y} = 1)$	$TP / (TP + FP)$	Precision
Negative Predictive Rate	$P(Y = 0 \hat{Y} = 0)$	$TN / (TN + FN)$	

Group Fairness

Demographic Parity (so called Statistical Parity)

The likelihood of a positive outcome should be the same across population groups.

- Mathematically, demographic parity requires that, for a protected attribute A ,

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$$

([Dwork et al., 2011](#))

- Variations:**

$$\frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)} \geq p \in [0, 1]$$

$$|P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1)| \leq \varepsilon \in [0, 1]$$

- The Four-Fifths rule:** if the selection rate for a certain group is less than 80% of that of the group with the highest selection rate, there is adverse impact on that group.

Equal Opportunity

The people who actually qualify for an opportunity are equally likely to get it regardless of their group membership.

- Mathematically, equal opportunity requires equal true positive rates (i.e., recall) across groups.

$$P(\hat{Y} = 1 \mid A = 0, Y = 1) = P(\hat{Y} = 1 \mid A = 1, Y = 1)$$

- Other related fairness definitions:**
 - Equalized odds** requires equal true positive rates and true negative rates across groups.
 - Conditional statistical parity:** Let L be a set of legitimate factors

$$P(\hat{Y} = 1 \mid A = 0, L = 1) = P(\hat{Y} = 1 \mid A = 1, L = 1)$$

Predictive Parity

The people who get the opportunity have an equal chance of being truly qualified regardless of their group membership.

- Mathematically, predictive parity requires equal precisions and equal negative predictive values across groups.

$$P(Y = 1 \mid A = 0, \hat{Y} = 1) = P(Y = 1 \mid A = 1, \hat{Y} = 1) \text{ \underline{AND} } P(Y = 0 \mid A = 0, \hat{Y} = 0) = P(Y = 0 \mid A = 1, \hat{Y} = 0)$$

- Related fairness definitions:**
 - Test fairness:** Let R be a predicted score (to be converted to \hat{Y} using a threshold). R is test fair if and only if $P(Y = 1 \mid A = 0, R = r) = P(Y = 1 \mid A = 1, R = r)$.
 - Well calibration** if and only if $P(Y = 1 \mid A = 0, R = r) = P(Y = 1 \mid A = 1, R = r) = r$.

The Impossibility Theorem ([Kleinberg et al., 2016](#))

Any two of the following three fairness definitions are **mutually exclusive**:

- | | |
|------------------------|--------------------------|
| (1) Demographic parity | $\hat{Y} \perp A$ |
| (2) Equalized odds | $\hat{Y} \perp A \mid Y$ |
| (3) Predictive parity | $Y \perp A \mid \hat{Y}$ |

except in the following two situations

- | | |
|----------------------|---|
| - Equal base rates | $P(Y = 1 \mid A = 0) = P(Y = 1 \mid A = 1)$ or |
| - Perfect prediction | $P(\hat{Y} = Y \mid A = 0) = P(\hat{Y} = Y \mid A = 1) = 1$ |

- So, we need to choose which definition we should adopt in our project.

Other Group Fairness Definitions

- Accuracy parity
- Error rate balance
 - False positive rate
 - False negative rate
- Predicted score balance
 - Positive class
 - Negative class

$$P(\hat{Y} = Y \mid A = 0) = P(\hat{Y} = Y \mid A = 1)$$

$$P(\hat{Y} = 1 \mid Y = 0, A = 0) = P(\hat{Y} = 1 \mid Y = 0, A = 1)$$

$$P(\hat{Y} = 0 \mid Y = 1, A = 0) = P(\hat{Y} = 0 \mid Y = 1, A = 1)$$

$$E(R \mid Y = 1, A = 0) = E(R \mid Y = 1, A = 1)$$

$$E(R \mid Y = 0, A = 0) = E(R \mid Y = 0, A = 1)$$

Subgroup Fairness

Subgroup Fairness

- Subgroup fairness is similar to group fairness but we consider a combination of more than one protected attribute.
- For example, given two binary protected attributes, A and B, equal opportunity is satisfied when

$$\begin{aligned} & P(\hat{Y} = 1 \mid A = 0, B = 0, Y = 1) \\ &= P(\hat{Y} = 1 \mid A = 0, B = 1, Y = 1) \\ &= P(\hat{Y} = 1 \mid A = 1, B = 0, Y = 1) \\ &= P(\hat{Y} = 1 \mid A = 1, B = 1, Y = 1) \end{aligned}$$

Individual Fairness

Individual Fairness (Definitions by [Kusner et al., 2017](#))

- **Fairness through unawareness**

An algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process.

- **Fairness through awareness**

An algorithm is fair if it gives similar predictions to similar individuals.

- This requires a carefully designed distance metric or similarity function.
- Advanced: Less qualified individuals should not be favored over more qualified individuals.

Counterfactual Fairness ([Kusner et al., 2017](#))

An algorithm is **counterfactually fair** if changing the values of any protected attributes does not affect the prediction.

- Mathematically, an algorithm is **counterfactually fair** if for any input $X = x$ and protected attribute $A = a$,

$$P(\hat{Y}_{A \leftarrow a} = y \mid X, A = a) = P(\hat{Y}_{A \leftarrow a'} = y \mid X, A = a)$$

for all y and for any value a' attainable by A .

From fairness definitions to evaluation metrics

- Demographic parity

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1) \longrightarrow m = |P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1)|$$

- False positive rate balance

$$P(\hat{Y} = 1 \mid Y = 0, A = 0) = P(\hat{Y} = 1 \mid Y = 0, A = 1) \longrightarrow m = |P(\hat{Y} = 1 \mid Y = 0, A = 0) - P(\hat{Y} = 1 \mid Y = 0, A = 1)|$$
$$\longrightarrow m = \sum_{a \in A} |\text{FPR} - \text{FPR}_a|$$