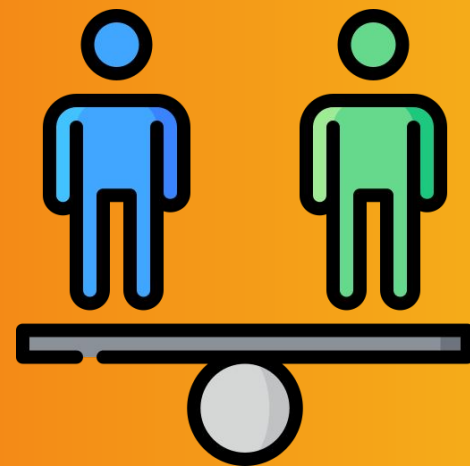


2021-22 Term 1

IS457: Fairness in Socio-technical Systems

Week 11 - Fairness mechanisms

KWAK Haewoon



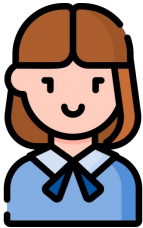
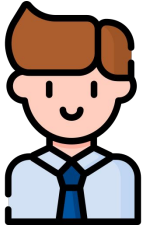
What are disparate impact and disparate treatment?

How can we formally define fairness?

How can we achieve fairness?

What are model cards?

ML for hiring - when individual-level info. is available



SMU SCIS	3.9	950	Exchange program in KR	Summer internship 3 months
SMU SCIS	3.9	950	Exchange program in KR	Summer internship 3 months

Is it acceptable if one applicant is recommended for hiring and the other is not?

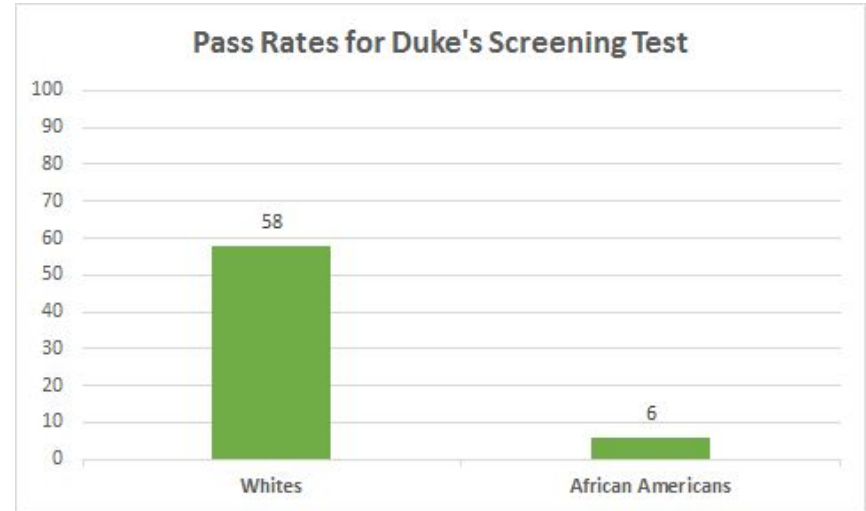
ML for hiring - when aggregated stats. are available

Imagine a situation that a selection rate of Blacks (among the Black applicants) is 10% and that of Whites (among the White applicants) is 50%.

Is it acceptable?

Griggs v. Duke Power Co. (1971)

“Congress has now provided that tests or criteria for employment or promotion may not provide equality of opportunity merely in the sense of the fabled offer of milk to the stork and the fox.”



Disparate impact vs. disparate treatment

Disparate treatment: when an algorithm provides different outputs for groups of people with the same (or similar) values of non-sensitive attributes (or features) but different values of sensitive attributes.

Disparate impact (“four-fifths rule / 80% rule”): when an algorithm provides outputs that benefit (hurt) a group of people sharing a value of sensitive attribute more frequently than other groups of people.

Questions

User Attributes		
Sensitive	Non-sensitive	
Gender	Clothing Bulge	Prox. Crime
Male 1	1	1
Male 2	1	0
Male 3	0	1
Female 1	1	1
Female 2	1	0
Female 3	0	0

Ground Truth (Has Weapon)
✓
✓
✗
✓
✗
✓

Classifier's Decision to Stop		
C ₁	C ₂	C ₃
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

Which algorithm is unfair due to disparate treatment?

Which algorithm is unfair due to disparate impact?

User Attributes		
Sensitive	Non-sensitive	
Gender	Clothing Bulge	Prox. Crime
Male 1	1	1
Male 2	1	0
Male 3	0	1
Female 1	1	1
Female 2	1	0
Female 3	0	0

Ground Truth (Has Weapon)
✓
✓
✗
✓
✗
✓

Classifier's Decision to Stop		
C ₁	C ₂	C ₃
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

Which algorithm is unfair due to disparate treatment? **C2**, **C3**

Which algorithm is unfair due to disparate impact? **C1**: 3/3 (Male) vs 2/3 (Female)

(Imaginary) COVID-19 test results

		COVID-19 test outcome	
Total population = 2000		Test outcome Positive	Test outcome Negative
Patients with COVID-19	Actual condition Positive	20	10
	Actual condition Negative	80	1890

When the **prediction** is **positive**,

True Positives (TP) are individuals for whom both the model prediction and actual outcome are positive labels.

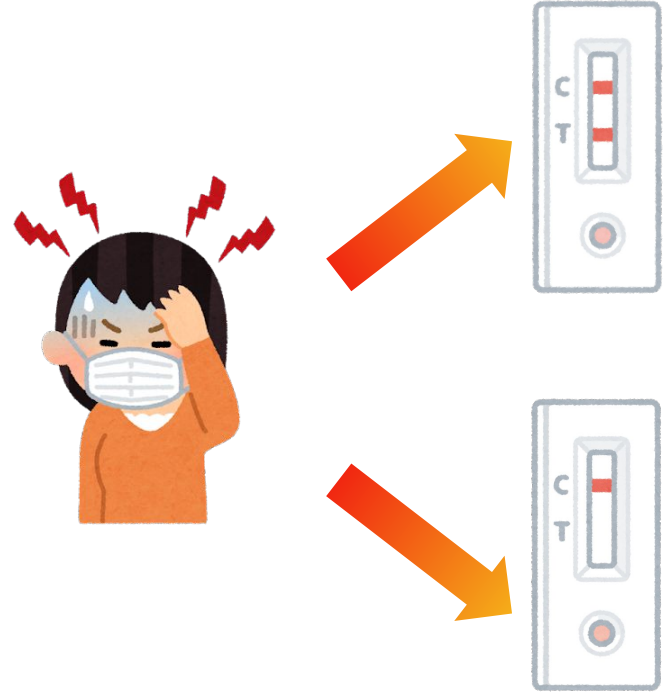
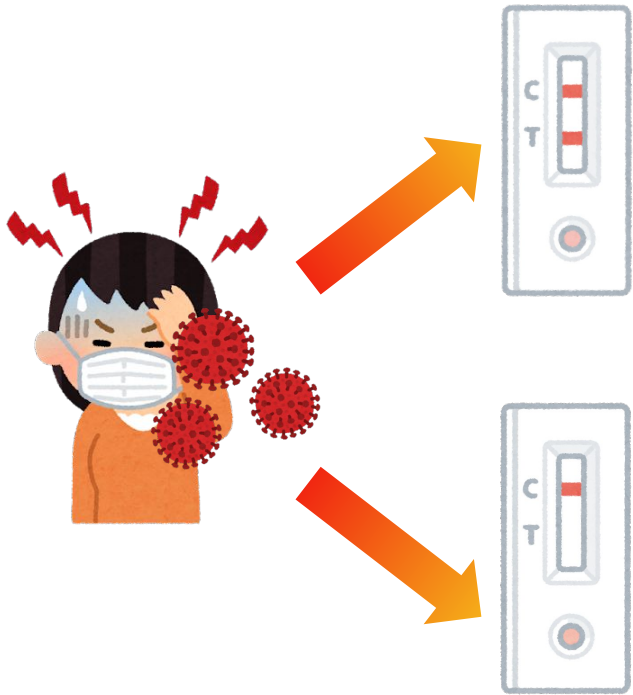
False Positives (FP) are individuals for whom both the model predicts a positive label, but the actual outcome is a negative label.

When the **prediction** is **negative**,

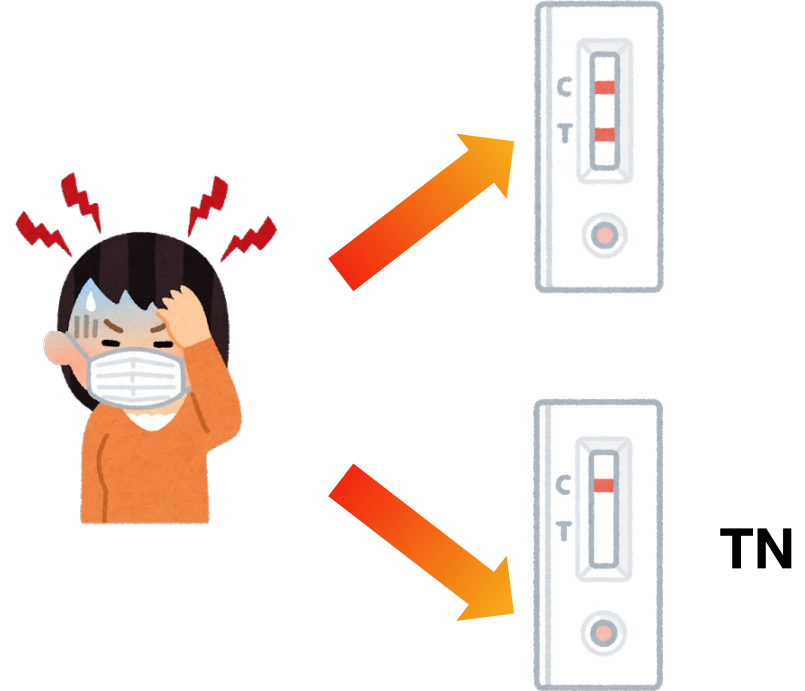
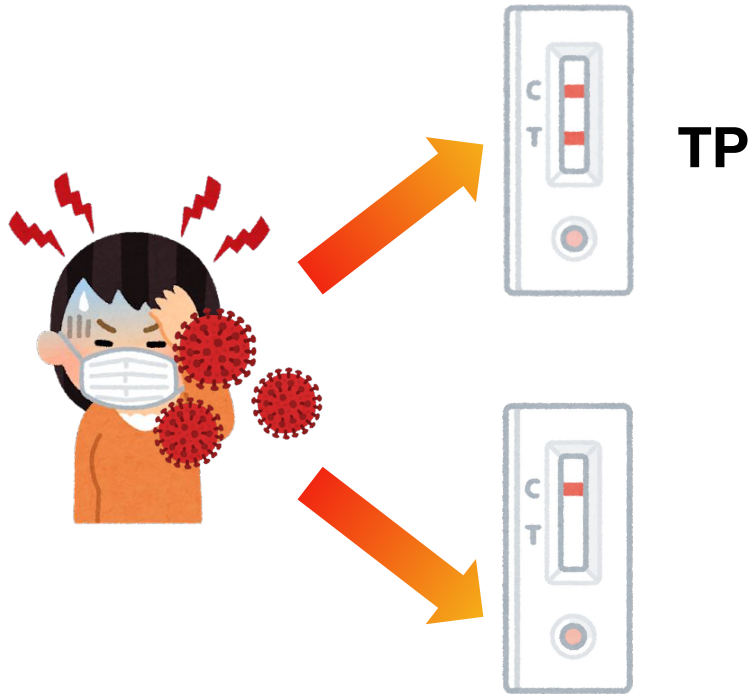
True Negatives (TN) are individuals for whom both the model prediction and actual outcome are negative labels.

False Negatives (FN) are individuals for whom both the model predicts a negative label, but the actual outcome is a positive label.

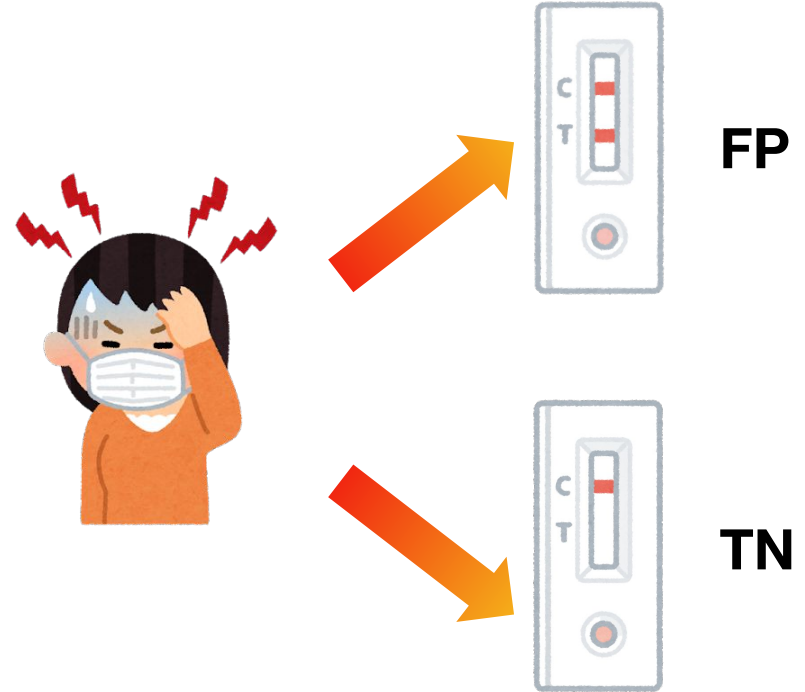
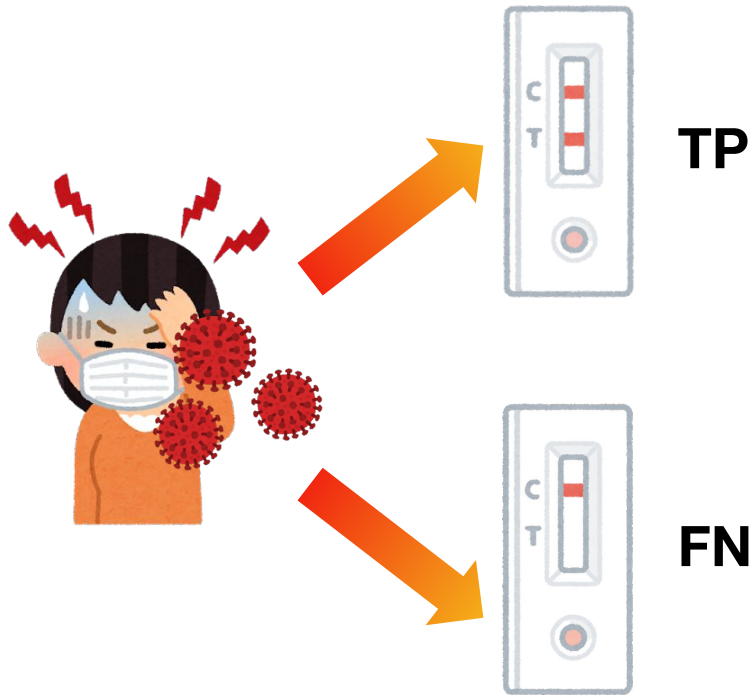
TP, FP, TN, FN in COVID-19 test



TP, FP, TN, FN in COVID-19 test



TP, FP, TN, FN in COVID-19 test



(Imaginary) COVID-19 test results

		COVID-19 test outcome	
		Test outcome Positive	Test outcome Negative
Patients with COVID-19	<i>pop</i> Total population = 2000		
	Actual condition Positive	<i>TP</i> True Positive 20	<i>FN</i> False Negative 10
	Actual condition Negative	<i>FP</i> False Positive 80	<i>TN</i> True Negative 1890

Accuracy and Prevalence

Accuracy (ACC)

$$= (TP + TN) / \text{pop} = (20 + 1890) / 2000$$

$$= 0.955$$

Prevalence (p)

$$= (\text{Actual condition positive}) / \text{pop}$$

$$= (TP + FN) / \text{pop} = (20 + 10) / 2000$$

$$= 0.015$$

		COVID-19 test outcome	
		Test outcome Positive	Test outcome Negative
Patients with COVID-19	<i>pop</i> Total population = 2000		
	Actual condition Positive	<i>TP</i> True Positive 20	<i>FN</i> False Negative 10
	Actual condition Negative	<i>FP</i> False Positive 80	<i>TN</i> True Negative 1890

TPR, FNR, FPR, TNR

		COVID-19 test outcome			
		Test outcome Positive	Test outcome Negative		
<i>pop</i> Total population = 2000					
Patients with COVID-19	Actual condition Positive	<i>TP</i> True Positive 20	<i>FN</i> False Negative 10	TPR True Positive Rate = $TP / (TP+FN)$	FNR False Negative Rate = $FN / (TP+FN)$
	Actual condition Negative	<i>FP</i> False Positive 80	<i>TN</i> True Negative 1890	FPR False Positive Rate = $FP / (FP+TN)$	TNR True Negative Rate = $TN / (FP+TN)$

PPV, FOR, FDR, NPV

	<i>pop</i> Total population = 2000	Test outcome Positive	Test outcome Negative
Patients with COVID-19	Actual condition Positive	<i>TP</i> True Positive 20	<i>FN</i> False Negative 10
	Actual condition Negative	<i>FP</i> False Positive 80	<i>TN</i> True Negative 1890
		PPV Positive predictive value = $TP / (TP+FP)$	FOR False omission rate = $FN / (FN+TN)$
		FDR False discovery rate = $FP / (TP+FP)$	NPV Negative predictive value = $TN / (FN+TN)$

In one figure

	<i>pop</i> Total population = 2000	Test outcome Positive	Test outcome Negative		
Patients with COVID-19	Actual condition Positive	<i>TP</i> True Positive 20	<i>FN</i> False Negative 10	TPR True Positive Rate = $TP / (TP+FN)$	FNR False Negative Rate = $FN / (TP+FN)$
	Actual condition Negative	<i>FP</i> False Positive 80	<i>TN</i> True Negative 1890	FPR False Positive Rate = $FP / (FP+TN)$	TNR True Negative Rate = $TN / (FP+TN)$
		PPV Positive predictive value = $TP / (TP+FP)$	FOR False omission rate = $FN / (FN+TN)$	※ Other names you might be heard Precision = PPV = 1 - FDR Recall = TPR = 1 - FNR	
		FDR False discovery rate = $FP / (TP+FP)$	NPV Negative predictive value = $TN / (FN+TN)$		

[Group activity] - Error metrics in context

[1 - Lending decisions]

Suppose you design an algorithm for a bank.

The algorithm decides if the bank should make the loan (1) or not (0).

[2 - Bail decisions]

Suppose you design an algorithm for criminal justice systems.

The algorithm decides if the defendant should be denied (1) or given (0) pretrial release.

Fill the [spreadsheet](#).

Is COVID-19 test fair across the gender?

Male population

	Total population	Test outcome Positive	Test outcome Negative		
Patients with COVID-19	Actual condition positive	True Positive	False Negative	TPR True Positive Rate $= TP / (TP+FN)$	FNR False Negative Rate $= FN / (TP+FN)$
	Actual condition negative	False Positive	True Negative	FPR False Positive Rate $= FP / (FP+TN)$	TNR True Negative Rate $= TN / (FP+TN)$
		PPV Positive predictive value $= TP / (TP+FP)$	FOR False omission rate $= FN / (FN+TN)$		
		FDR False discovery rate $= FP / (TP+FP)$	NPV Negative predictive value $= TN / (FN+TN)$		

Female population

	Total population	Test outcome Positive	Test outcome Negative		
Patients with COVID-19	Actual condition positive	True Positive	False Negative	TPR True Positive Rate $= TP / (TP+FN)$	FNR False Negative Rate $= FN / (TP+FN)$
	Actual condition negative	False Positive	True Negative	FPR False Positive Rate $= FP / (FP+TN)$	TNR True Negative Rate $= TN / (FP+TN)$
		PPV Positive predictive value $= TP / (TP+FP)$	FOR False omission rate $= FN / (FN+TN)$		
		FDR False discovery rate $= FP / (TP+FP)$	NPV Negative predictive value $= TN / (FN+TN)$		

Many of the error metrics in the previous page cannot be balanced across subgroups at the same time.

The most important aspect of measuring bias in your ML systems is understanding how “fairness” should be defined for your particular case. This requires:

- Considerations of the project’s goals
- Detailed discussions between the data scientists, decision makers, and those who will be affected by the application of the model.

While how fairness should be measured in a purely abstract manner is likely to be difficult, some guidelines can be provided.

When an algorithmic decision is assistive in nature (e.g., giving food subsidy):

Individuals may be harmed by failing to intervene on them when they have need, so you may care more about metrics that focus on false negatives.

When an algorithmic decision is punitive in nature (e.g., denying bail):

Individuals may be harmed by intervening on them in error so you may care more about metrics that focus on false positives.

Error metrics related to False negatives

Food subsidy programs:

	Give subsidy Positive	No subsidy Negative		
Need assistance	True Positive	False Negative	TPR True Positive Rate $= TP / (TP+FN)$	FNR False Negative Rate $= FN / (TP+FN)$
No assistance needed	False Positive	True Negative	FPR False Positive Rate $= FP / (FP+TN)$	TNR True Negative Rate $= TN / (FP+TN)$
	PPV Positive predictive value $= TP / (TP+FP)$	FOR False omission rate $= FN / (FN+TN)$		
	FDR False discovery rate $= FP / (TP+FP)$	NPV Negative predictive value $= TN / (FN+TN)$		

When different groups' FOR are the same.

In food subsidy program, FOR focuses the set of families not receiving the food subsidy. Such families will either be:

- True Negatives (those not in need of the assistance)
- False Negatives (those who need assistance but were missed)

FOR asks what fraction of this set fall into the latter category = $FN / (FN+TN)$

When different groups' FNR are the same.

In food subsidy program, FNR focuses instead on the set of people with need for the intervention. Such families will either be:

- True Positive (those who need assistance)
- False Negative (those who need assistance but were missed)

FNR asks what fraction of this set fall into the latter category = $FN / (TP + FN)$

Nuanced difference between FOR and FNR parity



School of
Computing and
Information Systems

FOR Parity: “If I were to choose a random family who do not receive the food subsidy from a given group, I would have the same chance of picking out one who need assistance across groups.”

FNR Parity: “...”

Nuanced difference between FOR and FNR parity



FOR Parity: “If I were to choose a random family who do not receive the food subsidy from a given group, I would have the same chance of picking out one who need assistance across groups.”

FNR Parity: “If I were to choose a random family with need for assistance from a given group, I would have the same chance of picking out one missed by the program across groups.”

Group A has 1000 total individuals. Among 200 who got subsidy, 80 are not in need of the assistance. Among the other 800, 40 actually needed help but were missed by the program.

Group B has 2000 total individuals. Among 1850 who got subsidy, 1400 are not in need of the assistance. Among the other 150, all actually needed help but were missed by the program.

Was the program fair in terms of FOR and FNR parity?



Answer - FNR parity is satisfied

	Give subsidy Positive	No subsidy Negative
Need assistance	120 True Positive	40 False Negative
No assistance needed	80 False Positive	760 True Negative

$$\text{FOR} = 40 / (40 + 760) = 0.05$$

$$\text{FNR} = 40 / (120 + 40) = 0.25$$

	Give subsidy Positive	No subsidy Negative
Need assistance	450 True Positive	150 False Negative
No assistance needed	1400 False Positive	0 True Negative

$$\text{FOR} = 150 / (150 + 0) = 1$$

$$\text{FNR} = 150 / (450 + 150) = 0.25$$

Error metrics related to False positives

Bail determination:

	Deny bail Positive	Grant bail Negative		
Who would Reoffend	True Positive	False Negative	TPR True Positive Rate $= TP / (TP+FN)$	FNR False Negative Rate $= FN / (TP+FN)$
Who would not reoffend	False Positive	True Negative	FPR False Positive Rate $= FP / (FP+TN)$	TNR True Negative Rate $= TN / (FP+TN)$
	PPV Positive predictive value $= TP / (TP+FP)$	FOR False omission rate $= FN / (FN+TN)$		
	FDR False discovery rate $= FP / (TP+FP)$	NPV Negative predictive value $= TN / (FN+TN)$		

When different groups' FDR are the same.

In bail decision, FDR focuses specifically on the people who are denied bail. Such people will either be:

- True Positives (those who would reoffend and were denied release)
- False Positives (those who would not reoffend but were kept in custody)

FDR asks what fraction of this set fall into the latter category = $FP / (TP + FP)$

When different groups' FPR are the same.

In bail decision, FPR focuses specifically on the people who will not reoffend.
Such people will either be:

- True Negatives (those who would not reoffend and were released)
- False Positives (those who would not reoffend but were kept in custody)

FPR asks what fraction of this set fall into the latter category = $FP / (TN + FP)$

Difference between FDR and FPR parity

FDR Parity: “If I were to choose a random people who are denied bail from a given group, I would have the same chance of picking out one who would not reoffend across groups.”

FPR Parity: “...”

Difference between FDR and FPR parity

FDR Parity: “If I were to choose a random people who are denied bail from a given group, I would have the same chance of picking out one who would not reoffend across groups.”

FPR Parity: “If I were to choose a random person who would not reoffend from a given group, I would have the same chance of picking out a person who were unnecessarily kept in custody across groups.”

Group A has 1000 total individuals, of whom 100 have been jailed with 10 wrongfully convicted. Suppose the other 900 are all guilty.

Group B has 3000 total individuals, of whom 300 have been jailed with 30 wrongfully convicted. Suppose the other 2700 are all innocent.

Were the verdicts fair in terms of FDR and FPR parity?

<https://www.wooclap.com/MKRGTP>



Answer - FDR parity is satisfied

Group A	Jailed Positive	Not Jailed Negative
Guilty	90 True Positive	900 False Negative
Not guilty	10 False Positive	0 True Negative

$$\text{FDR} = 10 / (10+90) = 0.1$$

$$\text{FPR} = 10 / (10+0) = 1$$

Group B	Jailed Positive	Not Jailed Negative
Guilty	270 True Positive	0 False Negative
Not guilty	30 False Positive	2700 True Negative

$$\text{FDR} = 30 / (30+270) = 0.1$$

$$\text{FPR} = 30 / (30+2700) = 0.01$$

More fairness criteria

Demographic parity

Overall accuracy equality

Equal opportunity

Demographic (statistical) parity

$P(\text{test} = \text{positive} \mid \text{group} = \text{male}) = P(\text{test} = \text{positive} \mid \text{group} = \text{female})$

$(\text{TP} + \text{FP}) / \text{pop}$ for male = $(\text{TP} + \text{FP}) / \text{pop}$ for female

→ Equal proportion of positive predictions in each group (No “disparate impact”)

	Algorithm output Positive	Algorithm output Negative		
Actual condition Positive	True Positive	False Negative	TPR True Positive Rate = $TP / (TP + FN)$	FNR False Negative Rate = $FN / (TP + FN)$
Actual condition Negative	False Positive	True Negative	FPR False Positive Rate = $FP / (FP + TN)$	TNR True Negative Rate = $TN / (FP + TN)$
PPV Positive predictive value = $TP / (TP + FP)$		FOR False omission rate = $FN / (FN + TN)$		
FDR False discovery rate = $FP / (TP + FP)$		NPV Negative predictive value = $TN / (FN + TN)$		

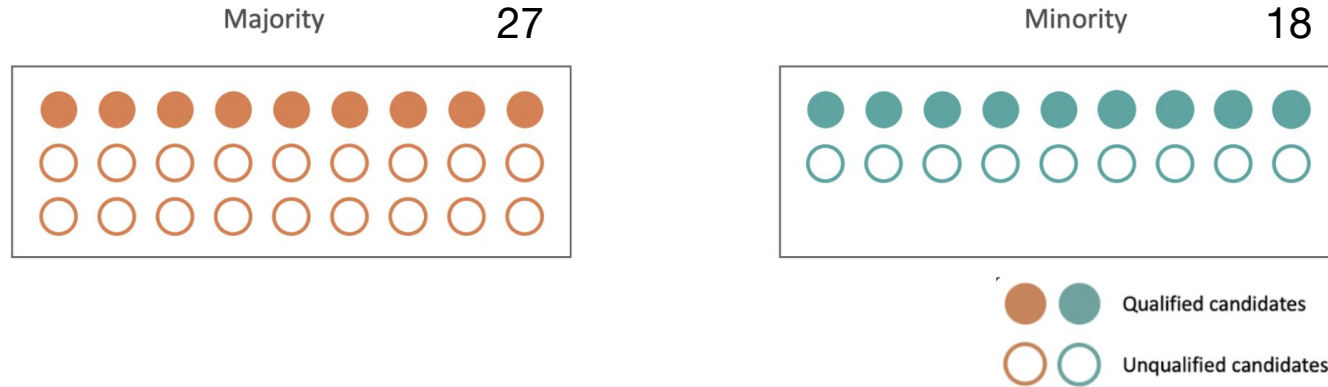
Demographic parity in context

(Bail) Same proportion of “bail denied” in each group (race).

(Lending) Same proportion of “loans granted” in each group.

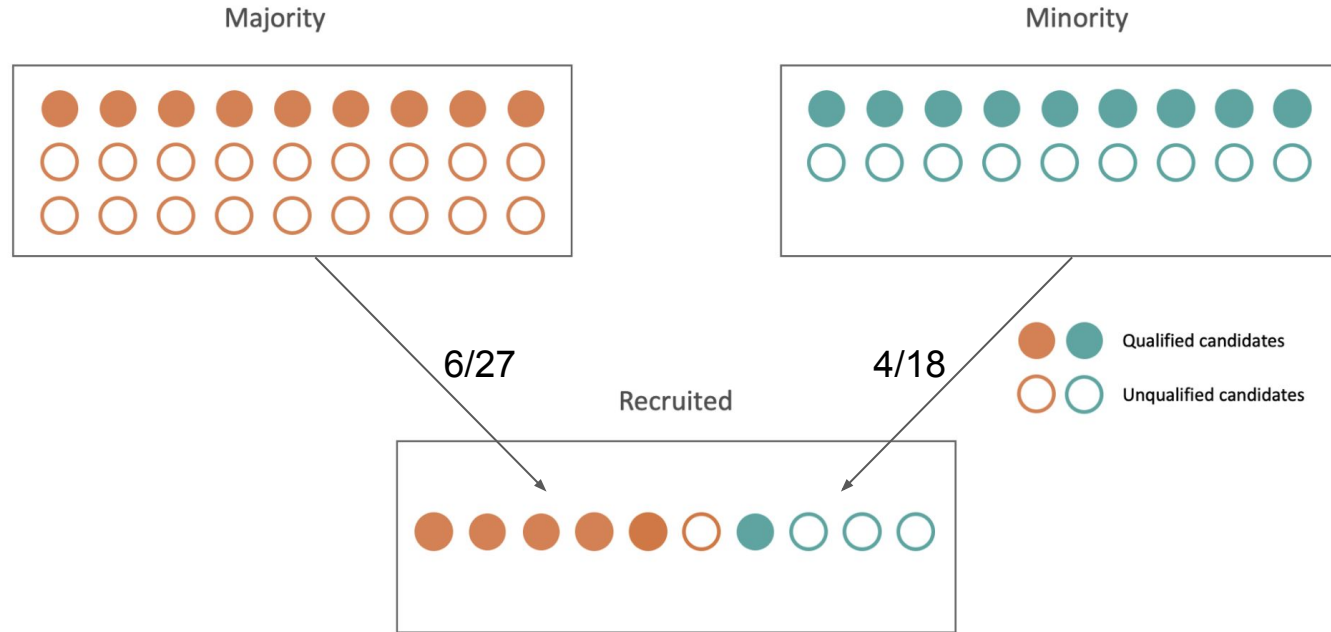
(Admissions) Same admission rate among each group.

Toy example - Hiring algorithm



If an algorithm recruits 15 candidates satisfying demographic parity, how many will be recommended from Majority and Minority groups?

Limitations (1) - when 10 are recruited

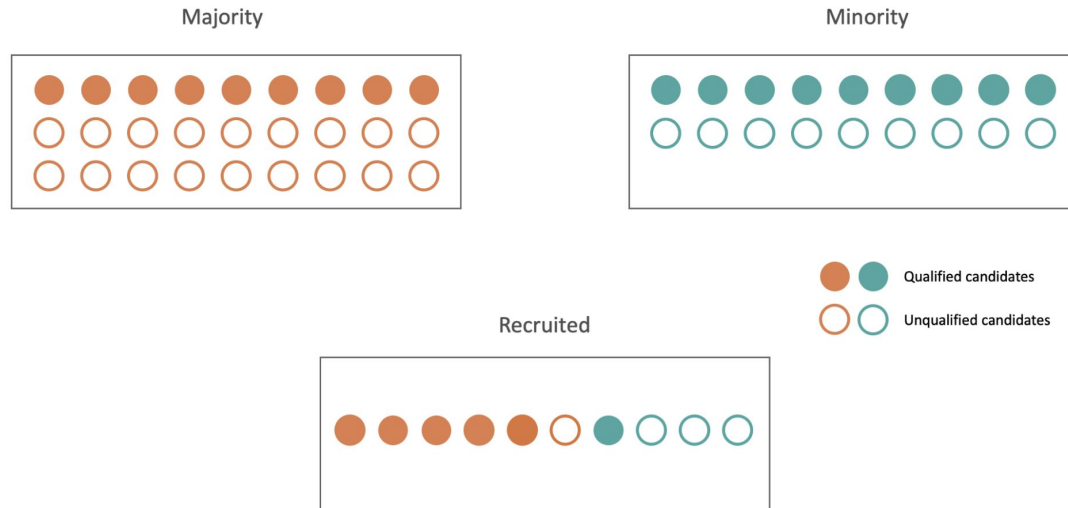


What are the potential harms of demographic parity?

Limitations (1)

Some qualified candidates could be treated unfairly.

Demographic parity does not guarantee the quality of predictions for each group.



Limitations (2)

If the prevalence is uneven across groups, a “perfect” classifier will not satisfy demographic parity.

	Algorithm output Positive	Algorithm output Negative		
Actual condition positive	True Positive	False Negative	TPR True Positive Rate $= TP / (TP+FN)$	FNR False Negative Rate $= FN / (TP+FN)$
Actual condition negative	False Positive	True Negative	FPR False Positive Rate $= FP / (FP+TN)$	TNR True Negative Rate $= TN / (FP+TN)$
	PPV Positive predictive value $= TP / (TP+FP)$	FOR False omission rate $= FN / (FN+TN)$		
	FDR False discovery rate $= FP / (TP+FP)$	NPV Negative predictive value $= TN / (FN+TN)$		

	Algorithm output Positive	Algorithm output Negative		
Actual condition positive	True Positive	False Negative	TPR True Positive Rate $= TP / (TP+FN)$	FNR False Negative Rate $= FN / (TP+FN)$
Actual condition negative	False Positive	True Negative	FPR False Positive Rate $= FP / (FP+TN)$	TNR True Negative Rate $= TN / (FP+TN)$
	PPV Positive predictive value $= TP / (TP+FP)$	FOR False omission rate $= FN / (FN+TN)$		
	FDR False discovery rate $= FP / (TP+FP)$	NPV Negative predictive value $= TN / (FN+TN)$		

Overall accuracy equality

$P(\text{test} = \text{actual} \mid \text{group}=\text{male}) = P(\text{test} = \text{actual} \mid \text{group} = \text{female})$

$(\text{TP}+\text{TN})/\text{pop}$ is the same for each group.

Error parity is sometimes computed.

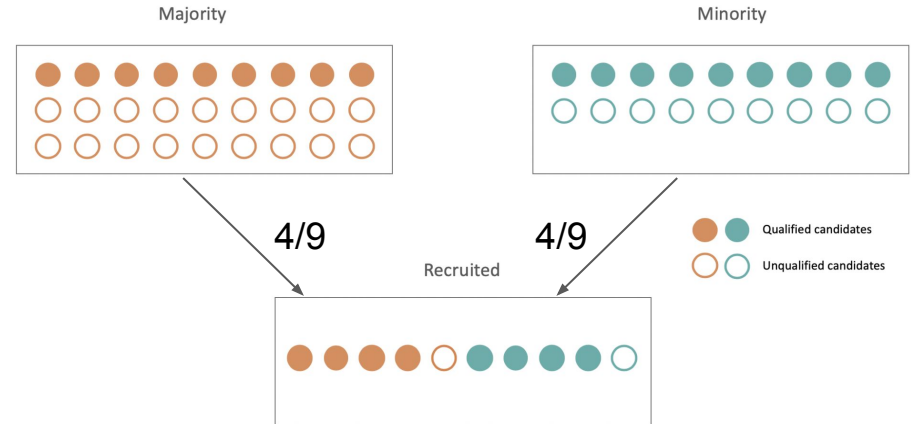
$(\text{FP}+\text{FN})/\text{pop}$ is the same for each group.

	Algorithm output Positive	Algorithm output Negative		
Actual condition Positive	True Positive	False Negative	TPR True Positive Rate = $\text{TP} / (\text{TP}+\text{FN})$	FNR False Negative Rate = $\text{FN} / (\text{TP}+\text{FN})$
Actual condition Negative	False Positive	True Negative	FPR False Positive Rate = $\text{FP} / (\text{FP}+\text{TN})$	TNR True Negative Rate = $\text{TN} / (\text{FP}+\text{TN})$
	PPV Positive predictive value = $\text{TP} / (\text{TP}+\text{FP})$	FOR False omission rate = $\text{FN} / (\text{FN}+\text{TN})$		
	FDR False discovery rate = $\text{FP} / (\text{TP}+\text{FP})$	NPV Negative predictive value = $\text{TN} / (\text{FN}+\text{TN})$		

$$P(\text{test} = \text{pos} \mid \text{actual} = \text{pos}, \text{group} = M) = P(\text{test} = \text{pos} \mid \text{actual} = \text{pos}, \text{group} = F)$$




Two groups have equal TPR.

	Algorithm output Positive	Algorithm output Negative
Actual condition positive	True Positive	False Negative
Actual condition negative	False Positive	True Negative
	TPR True Positive Rate = $TP / (TP + FN)$	FNR False Negative Rate = $FN / (TP + FN)$
	FPR False Positive Rate = $FP / (FP + TN)$	TNR True Negative Rate = $TN / (FP + TN)$
	PPV Positive predictive value = $TP / (TP + FP)$	FOR False omission rate = $FN / (FN + TN)$
	FDR False discovery rate = $FP / (TP + FP)$	NPV Negative predictive value = $TN / (FN + TN)$



Bias in criminal justice systems

In Week 2



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

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

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

May 23, 2016

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Finding 2: Racial disparities exist

In Week 2

In forecasting who would re-offend, the algorithm made mistakes with black and white defendant at roughly the same rate but:

- Falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.

	Algorithm output Positive	Algorithm output Negative		
Actual condition positive	True Positive	False Negative	TPR True Positive Rate $= TP / (TP+FN)$	FNR False Negative Rate $= FN / (TP+FN)$
Actual condition negative	False Positive	True Negative	FPR False Positive Rate $= FP / (FP+TN)$	TNR True Negative Rate $= TN / (FP+TN)$
	PPV Positive predictive value $= TP / (TP+FP)$	FOR False omission rate $= FN / (FN+TN)$		
	FDR False discovery rate $= FP / (TP+FP)$	NPV Negative predictive value $= TN / (FN+TN)$		

AA population

	Higher risk Positive	Lower risk Negative
Who would reoffend	True Positive 1369	False Negative 532
Who would not reoffend	False Positive 805	True Negative 990

White population

	Higher risk Positive	Lower risk Negative
Who would reoffend	True Positive 505	False Negative 461
Who would not reoffend	False Positive 349	True Negative 1139

FPR and FNR parity are violated.

AA:

$$\text{FPR} = \text{FP} / (\text{FP} + \text{TN}) = 805 / (805 + 990) = 0.45$$

$$\text{FNR} = \text{FN} / (\text{TP} + \text{FN}) = 532 / (1369 + 532) = 0.28$$

White:

$$\text{FPR} = \text{FP} / (\text{FP} + \text{TN}) = 349 / (349 + 1139) = 0.23$$

$$\text{FNR} = \text{FN} / (\text{TP} + \text{FN}) = 461 / (505 + 461) = 0.48$$

FDR parity is satisfied.

AA:
$$\text{FDR} = \text{FP} / (\text{TP} + \text{FP}) = 805 / (1369 + 805) = 0.37$$

White:
$$\text{FDR} = \text{FP} / (\text{TP} + \text{FP}) = 349 / (505 + 349) = 0.41$$

Who is right?

ProPublica: “If you’re a person who would not reoffend, the probability of being mislabeled as high risk is different based on your race. Thus, the algorithm is unfair.”

NorthPointe: “If the model labels you as high risk, the actual probability that you would not reoffend is not different based on your race. Thus, the algorithm is fair.”

VERNON PRATER	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	HIGH RISK 8

(In the presence of different base rates,) it would be impossible for a model to satisfy both definitions of fairness at the same time (“impossibility theorem”).

The **researchers** developing a model, the **policymakers** deciding to put it into practice, and the **users** making decisions based upon the model should understand and explore the options for measuring fairness as well as the trade-offs involved in making that choice.



Collect more training data: usually expensive, sometimes impossible.

Auditing model

Model selection

Auditing existing models helps understand whether that process is yielding equitable results.

The existing model could be any set of decisions and outcomes.

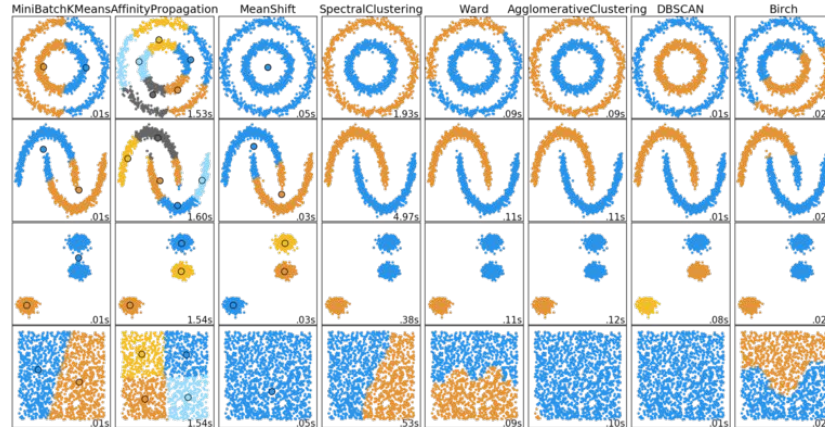
It will generally be useful to make measures of equity in any existing processes.

- It will help to decide whether a new model might improve, degrade, or leave unchanged the fairness of the existing system.

Model selection (1)

Many models can be developed for a given problem, making the task of selecting a specific model an important step in the process of model development.

Model selection can naturally be extended to incorporate fairness metrics.



No one-size-fits-all metric works in all contexts. We need to ask:

- If many models perform similarly on overall evaluation metrics of interest (e.g., accuracy), how do they vary in terms of fairness?
- How much “cost” in terms of performance do you have to pay to reach various levels of fairness?
- Which model performs best on each of fairness metrics?

In most cases, model selection will yield a number of options for a final model.

Unlike model selection based on performance metrics alone, the final choice between these will generally involve a judgment call that reflects the project's dual goals of balancing accuracy and fairness.

The final choice of model is best treated as a discussion between the data scientists and stakeholders in the same manner as choosing how to define fairness in the first place.

Model cards for model reporting

Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions.

Model cards clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited.

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.

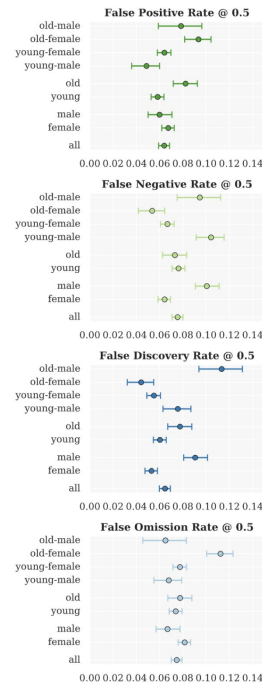
Ethical Considerations

- 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



Model developers can compare the model's results to other models in the same space, and make decisions about training their own system.

Software developers working on products that use the model's predictions can inform their design and implementation decisions.

Policymakers can understand how a machine learning system may fail or succeed in ways that impact people.

Organizations can inform decisions about adopting technology that incorporates machine learning.

Impacted individuals who may experience effects from a model can better understand how it works or use information in the card to pursue remedies.

1. Model details

Basic information about the model.

- Person or organization developing model
- Model date
- Model version
- Model type
- Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
- Paper or other resource for more information
- Citation details
- License
- Where to send questions or comments about the model

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.

Metrics

- 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].
- 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.

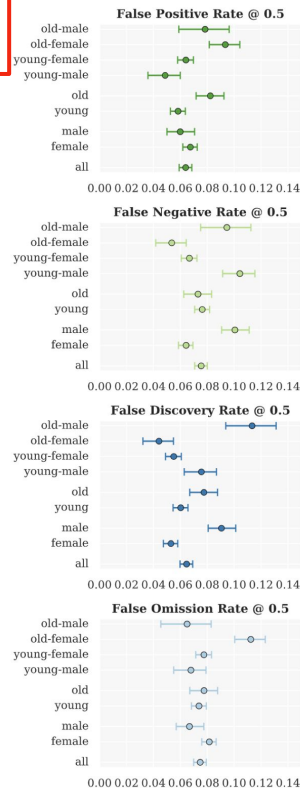
Ethical Considerations

- 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



2. Intended use

Use cases that were envisioned during development.

- Primary intended uses
- Primary intended users
- Out-of-scope use cases (e.g., for use on black-and-white images only; not for use on text examples shorter than 140 chars)

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.

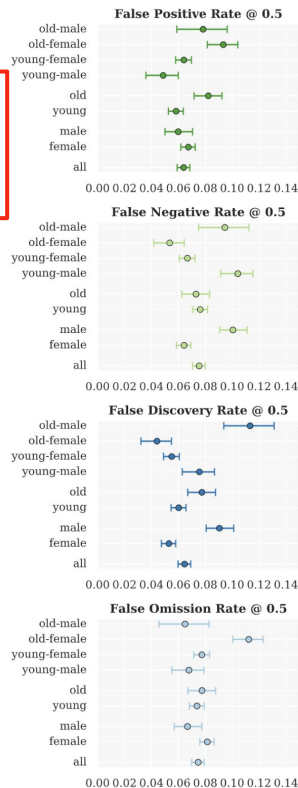
Ethical Considerations

- 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



3. Factors

Factors could include demographic or phenotypic groups, environmental conditions, and technical attributes

- Relevant factors: What are foreseeable salient factors for which model performance may vary, and how were these determined?
- Evaluation factors: Which factors are being reported, and why were these chosen?

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.

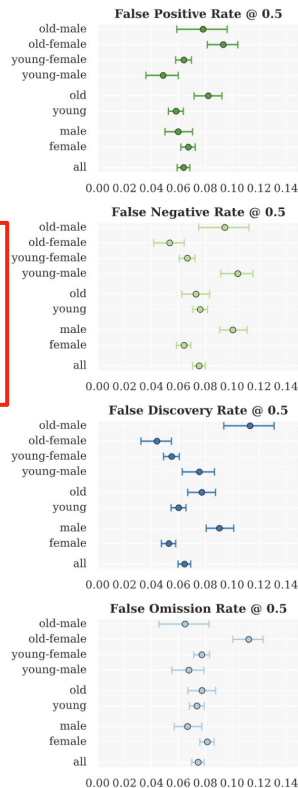
Ethical Considerations

- 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



4. Metrics

Metrics should be chosen to reflect potential real-world impacts of the model.

- Model performance measures
- Decision thresholds
- Variation approaches

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.

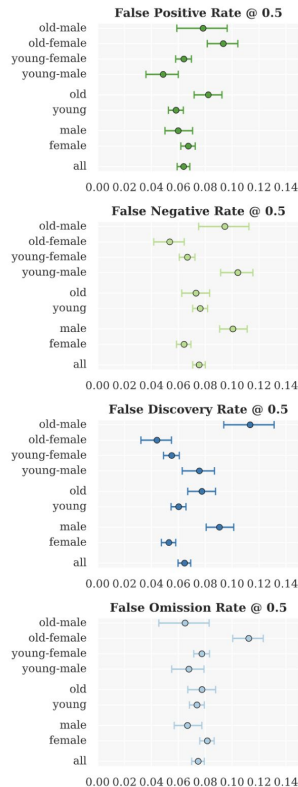
Ethical Considerations

- 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



5. Evaluation data

Details on the dataset(s) used for the quantitative analyses in the card.

- Datasets: This should include datasets that are publicly available for third-party use.
- Motivation
- Preprocessing

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

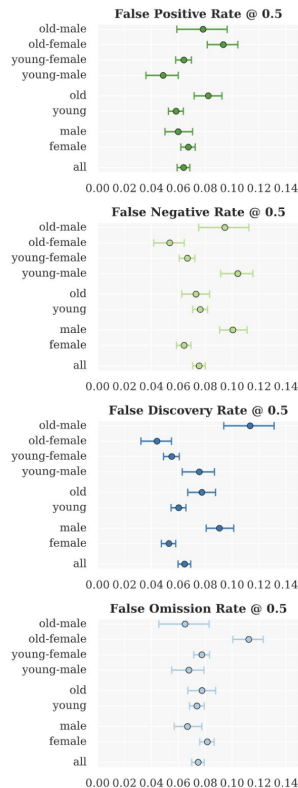
Ethical Considerations

- 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



6. Training data

May not be possible to provide in practice (e.g., the data may be proprietary, or require a non-disclosure agreement).

If so, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.



7. Quantitative analyses

Results of evaluating the model according to the chosen metrics

- Unitary results: How did the model perform with respect to each factor?
- Intersectional results: How did the model perform with respect to the intersection of evaluated factors?

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.

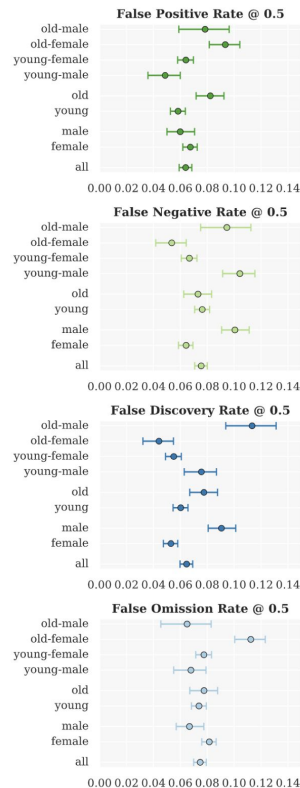
Ethical Considerations

- 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



8. Ethical considerations

Demonstrate the ethical considerations that went into model development, surfacing ethical challenges and solutions to stakeholders.

- Data: Does the model use any sensitive data?
- Risk and harms: What risks may be present in model usage?

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

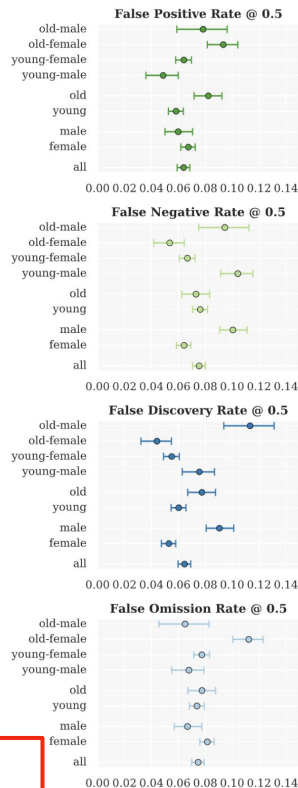
Ethical Considerations

- 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



9. Caveats and recommendations

Additional concerns that were not covered in the previous sections.

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

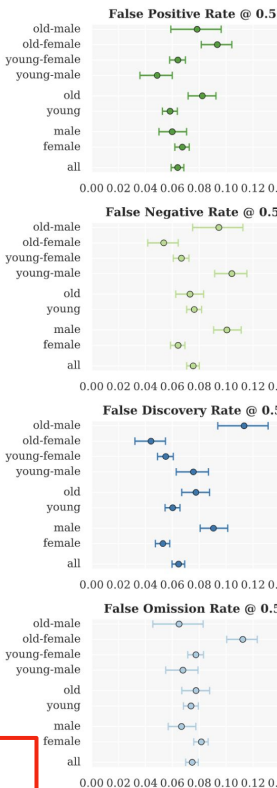
Ethical Considerations

- 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



HuggingFace (e.g., DistilBERT <https://huggingface.co/distilbert-base-uncased>)

When sharing a model, what should I add to my model card?

The model card should describe:

- the model
- its intended uses & potential limitations, including bias and ethical considerations as detailed in [Mitchell, 2018](#)
- the training params and experimental info (you can embed or link to an experiment tracking platform for reference)
- which datasets did you train on and your eval results

Reflection

<https://smu.sg/IS457r11>