

# FAIRNESS: DEFINITIONS AND MEASUREMENTS

Eunsuk Kang

Required reading: Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach.  
"Improving fairness in machine learning systems: What do industry practitioners need?" In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2019.

# LEARNING GOALS

- Understand different definitions of fairness
- Discuss methods for measuring fairness
- Consider fairness throughout an ML lifecycle

# FAIRNESS: DEFINITIONS

# FAIRNESS IS STILL AN ACTIVELY STUDIED & DISPUTED CONCEPT!



Source: Mortiz Hardt, <https://fairmlclass.github.io/>

# FAIRNESS: DEFINITIONS

- Anti-classification (fairness through blindness)
- Independence (group fairness)
- Separation (equalized odds)
- ...and numerous others!

# ANTI-CLASSIFICATION



- Also called *fairness through blindness*
- Ignore/eliminate sensitive attributes from dataset
- Example: Remove gender or race from a credit card scoring system
- **Q. Advantages and limitations?**

# RECALL: PROXIES

*Features correlate with protected attributes*



# RECALL: NOT ALL DISCRIMINATION IS HARMFUL



FEDERAL TRADE COMMISSION

Mortgage discrimination is against the law.



- Loan lending: Gender discrimination is illegal.
- Medical diagnosis: Gender-specific diagnosis may be desirable.
- Discrimination is a **domain-specific** concept!

Other examples?



# ANTI-CLASSIFICATION



- Ignore/eliminate sensitive attributes from dataset
- Limitations
  - Sensitive attributes may be correlated with other features
  - Some ML tasks need sensitive attributes (e.g., medical diagnosis)

# TESTING ANTI-CLASSIFICATION

How do we test that an ML model achieves anti-classification?

# TESTING ANTI-CLASSIFICATION

Straightforward invariant for classifier  $f$  and protected attribute  $p$ :

$$\forall x. f(x[p \leftarrow 0]) = f(x[p \leftarrow 1])$$

*(does not account for correlated attributes)*

Test with random input data or on any test data

Any single inconsistency shows that the protected attribute was used. Can also report percentage of inconsistencies.

See for example: Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou. "[Fairness testing: testing software for discrimination](#)." In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, pp. 498-510. 2017.

# NOTATIONS

- $X$ : Feature set (e.g., age, race, education, region, income, etc.,)
- $A \in X$ : Sensitive attribute (e.g., gender)
- $R$ : Regression score (e.g., predicted likelihood of loan default)
- $Y'$ : Classifier output
  - $Y' = 1$  if and only if  $R > T$  for some threshold  $T$
  - e.g., Deny the loan ( $Y' = 1$ ) if the likelihood of default  $> 30\%$
- $Y$ : Target variable being predicted ( $Y = 1$  if the person actually defaults on loan)

Setting classification thresholds: Loan lending example

# INDEPENDENCE

$$P[Y' = 1 | A = a] = P[Y' = 1 | A = b]$$

- Also called *group fairness* or *demographic parity*
- Mathematically,  $Y' \perp A$ 
  - Prediction ( $Y'$ ) must be independent of the sensitive attribute ( $A$ )
- Examples:
  - The predicted rate of recidivism is the same across all races
  - Both women and men have the equal probability of being promoted
  - i.e.,  $P[\text{promote} = 1 | \text{gender} = M] = P[\text{promote} = 1 | \text{gender} = F]$

# INDEPENDENCE

# INDEPENDENCE

- Q. What are limitations of independence?

# INDEPENDENCE

- Q. What are limitations of independence?
  - Ignores possible correlation between  $Y$  and  $A$ 
    - Rules out perfect predictor  $Y' = Y$  when  $Y$  &  $A$  are correlated



# INDEPENDENCE

- Q. What are limitations of independence?
  - Ignores possible correlation between  $Y$  and  $A$ 
    - Rules out perfect predictor  $Y' = Y$  when  $Y$  &  $A$  are correlated
  - Permits abuse and laziness: Can be satisfied by randomly assigning a positive outcome ( $Y' = 1$ ) to protected groups
    - e.g., Randomly promote people (regardless of their job performance) to match the rate across all groups

# RECALL: EQUALITY VS EQUITY



# CALIBRATION TO ACHIEVE INDEPENDENCE

Select different thresholds for different groups to achieve prediction parity:

$$P[R > t_0 | A = 0] = P[R > t_1 | A = 1]$$

Lowers bar for some groups -- equity, not equality

# TESTING INDEPENDENCE

# TESTING INDEPENDENCE

- Separate validation/telemetry data by protected attribute

# TESTING INDEPENDENCE

- Separate validation/telemetry data by protected attribute
  - Or generate realistic test data, e.g. from probability distribution of population

# TESTING INDEPENDENCE

- Separate validation/telemetry data by protected attribute
  - Or generate realistic test data, e.g. from probability distribution of population
- Separately measure rate of positive predictions

# TESTING INDEPENDENCE

- Separate validation/telemetry data by protected attribute
  - Or generate realistic test data, e.g. from probability distribution of population
- Separately measure rate of positive predictions
- Report issue if rate differs beyond  $\epsilon$  across groups



# SEPARATION

$$P[Y' = 1 \mid Y = 0, A = a] = P[Y' = 1 \mid Y = 0, A = b]$$

$$P[Y' = 0 \mid Y = 1, A = a] = P[Y' = 0 \mid Y = 1, A = b]$$

- Also called *equalized odds*
- $Y' \perp A \mid Y$ 
  - Prediction must be independent of the sensitive attribute *conditional* on the target variable

# REVIEW: CONFUSION MATRIX

		Actual value	
		$Y = 1$	$Y = 0$
Predicted value	$Y' = 1$	True Positive Rate $P[Y' = 1 \mid Y = 1]$	False Positive Rate $P[Y' = 1 \mid Y = 0]$
	$Y' = 0$	False Negative Rate $P[Y' = 0 \mid Y = 1]$	True Negative Rate $P[Y' = 0 \mid Y = 0]$

Can we explain separation in terms of model errors?

$$P[Y' = 1 \mid Y = 0, A = a] = P[Y' = 1 \mid Y = 0, A = b]$$

$$P[Y' = 0 \mid Y = 1, A = a] = P[Y' = 0 \mid Y = 1, A = b]$$



# SEPARATION

$$P[Y' = 1 \mid Y = 0, A = a] = P[Y' = 1 \mid Y = 0, A = b] \text{ (FPR parity)}$$

$$P[Y' = 0 \mid Y = 1, A = a] = P[Y' = 0 \mid Y = 1, A = b] \text{ (FNR parity)}$$

- $Y' \perp A \mid Y$ 
  - Prediction must be independent of the sensitive attribute *conditional* on the target variable

# SEPARATION

$$P[Y' = 1 \mid Y = 0, A = a] = P[Y' = 1 \mid Y = 0, A = b] \text{ (FPR parity)}$$

$$P[Y' = 0 \mid Y = 1, A = a] = P[Y' = 0 \mid Y = 1, A = b] \text{ (FNR parity)}$$

- $Y' \perp A \mid Y$ 
  - Prediction must be independent of the sensitive attribute *conditional* on the target variable
- i.e., All groups are susceptible to the same false positive/negative rates

# SEPARATION

$$P[Y' = 1 \mid Y = 0, A = a] = P[Y' = 1 \mid Y = 0, A = b] \text{ (FPR parity)}$$

$$P[Y' = 0 \mid Y = 1, A = a] = P[Y' = 0 \mid Y = 1, A = b] \text{ (FNR parity)}$$

- $Y' \perp A \mid Y$ 
  - Prediction must be independent of the sensitive attribute *conditional* on the target variable
- i.e., All groups are susceptible to the same false positive/negative rates
- Example: Promotion
  - Y': Promotion decision, A: Gender of applicant: Y: Actual job performance
  - Separation w/ FNR: Probability of being incorrectly denied promotion is equal across both male & female employees

# TESTING SEPARATION

- Generate separate validation sets for each group
- Separate validation/telemetry data by protected attribute
  - Or generate *realistic* test data, e.g. from probability distribution of population
- Separately measure false positive and false negative rates

# CASE STUDY: CANCER DIAGNOSIS





# EXERCISE: CANCER DIAGNOSIS

## Overall Results

True positives (TPs): 16

False positives (FPs): 21

False negatives (FNs): 9

True negatives (TNs): 954

## Male Patient Results

True positives  
(TPs): 3

False positives  
(FPs): 16

False negatives  
(FNs): 7

True negatives  
(TNs): 474

## Female Patient Results

True positives  
(TPs): 13

False positives  
(FPs): 5

False negatives  
(FNs): 2

True negatives  
(TNs): 480

- 1000 data samples (500 male & 500 female patients)
- Does the model achieve independence? Separation w/ FPR or FNR?
- What can we conclude about the model & its usage?

# REVIEW OF CRITERIA SO FAR:

*Recidivism scenario: Should a person be  
detained?*

- Anti-classification: ?
- Independence: ?
- Separation: ?





# REVIEW OF CRITERIA SO FAR:

*Recidivism scenario: Should a defendant be detained?*

- Anti-classification: Race and gender should not be considered for the decision at all
- Independence: Detention rates should be equal across gender and race groups
- Separation: Among defendants who would not have gone on to commit a violent crime if released, detention rates are equal across gender and race groups

# ACHIEVING FAIRNESS CRITERIA

# CAN WE ACHIEVE FAIRNESS DURING THE LEARNING PROCESS?

- Data acquisition:
  - Collect additional data if performance is poor on some groups
- Pre-processing:
  - Clean the dataset to reduce correlation between the feature set and sensitive attributes
- Training time constraint
  - ML is a constraint optimization problem (i.e., minimize errors)
  - Impose additional parity constraint into ML optimization process (as part of the loss function)
- Post-processing
  - Adjust thresholds to achieve a desired fairness metric
- (Still active area of research! Many new techniques published each year)

*Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints*, Cotter et al., (2018).

# TRADE-OFFS: ACCURACY VS FAIRNESS



- In general, accuracy is at odds with fairness
  - e.g., Impossible to achieve perfect accuracy ( $R = Y$ ) while ensuring independence
- Determine how much compromise in accuracy or fairness is acceptable to your stakeholders





# BUILDING FAIR ML SYSTEMS

# FAIRNESS MUST BE CONSIDERED THROUGHOUT THE ML LIFECYCLE!





# PRACTITIONER CHALLENGES

- Fairness is a system-level property
  - consider goals, user interaction design, data collection, monitoring, model interaction (properties of a single model may not matter much)
- Fairness-aware data collection, fairness testing for training data
- Identifying blind spots
  - Proactive vs reactive
  - Team bias and (domain-specific) checklists
- Fairness auditing processes and tools
- Diagnosis and debugging (outlier or systemic problem? causes?)
- Guiding interventions (adjust goals? more data? side effects? chasing mistakes? redesign?)
- Assessing human bias of humans in the loop

Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "[Improving fairness in machine learning systems: What do industry practitioners need?](#)" In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2019.

# **REQUIREMENTS ENGINEERING FOR FAIRNESS**

# RECALL: MACHINE VS WORLD



- No ML/AI lives in vacuum; every system is deployed as part of the world
- A requirement describes a desired state of the world (i.e., environment)
- Machine (software) is *created* to manipulate the environment into this state

# REQUIREMENTS FOR FAIR ML SYSTEMS

- Identify requirements (REQ) over the environment
  - What types of harm can be caused by biased decisions?
  - Who are stakeholders? Which population groups can be harmed?
  - Are we trying to achieve equality vs. equity?
  - What are legal requirements to consider?
- Define the interface between the environment & machine (ML)
  - What data will be sensed/measured by AI? Potential biases?
  - What types of decisions will the system make? Punitive or assistive?
- Identify the environmental assumptions (ENV)
  - Adversarial? Misuse? Unfair (dis-)advantages?
  - Population distributions?
- Devise machine specifications (SPEC) that are sufficient to establish REQ
  - What type of fairness definition is appropriate?

# "FOUR-FIFTH RULE" (OR "80% RULE")

$$(P[R = 1 | A = a]) / (P[R = 1 | A = b]) \geq 0.8$$

- Selection rate for a protected group (e.g.,  $A = a$ )  $< 80\%$  of highest rate  $\Rightarrow$  selection procedure considered as having "adverse impact"
- Guideline adopted by Federal agencies (Department of Justice, Equal Employment Opportunity Commission, etc.,) in 1978
- If violated, must justify business necessity (i.e., the selection procedure is essential to the safe & efficient operation)
- Example: Hiring
  - 50% of male applicants vs 20% female applicants hired ( $0.2/0.5 = 0.4$ )
  - Is there a business justification for hiring men at a higher rate?



## EXAMPLE: LOAN APPLICATION

**LOAN APPLICATION**

**Personal Information**

Name (Last) PUBLIC (First) JON (Middle Initial) J  
 Address (Mailing Address) 12345 MAIN STREET (City) ANCHORAGE (State) AK (Zip) 99501  
 E-Mail Address JQPJQPJQP@EXAMPLE.COM (Home Telephone) 1111 1111 (Other Telephone) 222 222 2222

**Services needed**

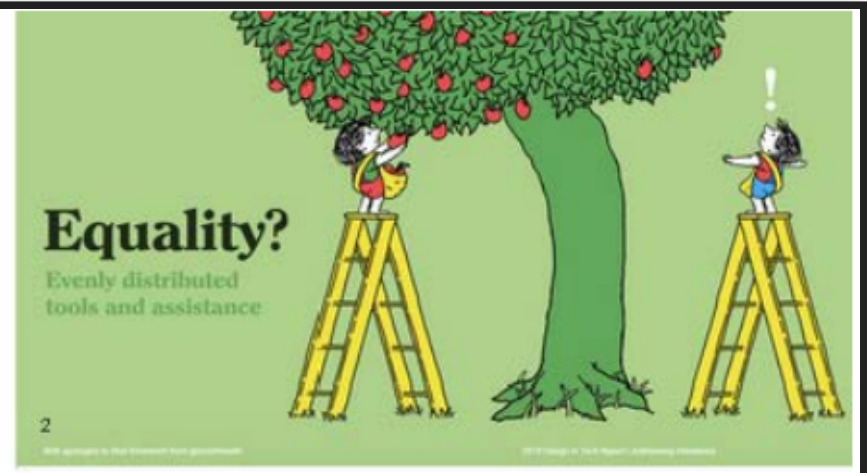
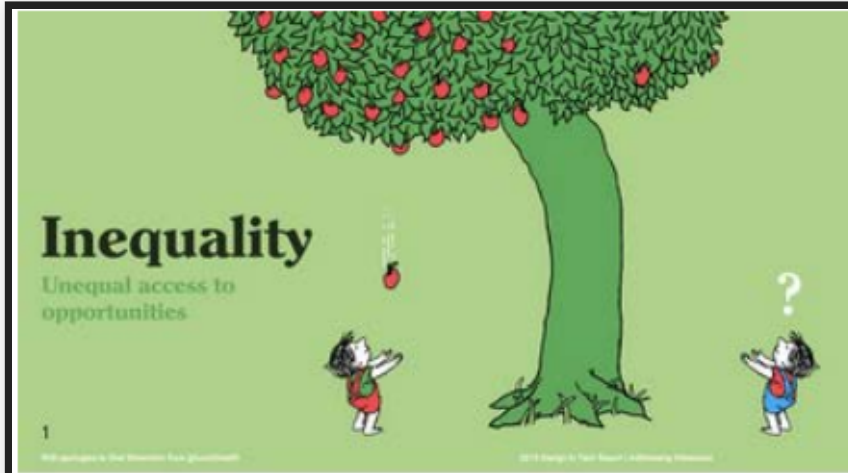
UNDER REVIEW

**Current Income**

High School Graduate Or General Education (GED) Test Passed? Yes No  
 Highest grade completed  
 Military (Most recent first) Credits Earned  
 Quarterly or Other (Specify) Graduate Major or Subject

- Who are the stakeholders?
- Types of harm?
- Legal & policy considerations?

# RECALL: EQUALITY VS EQUITY



# TYPE OF DECISION & POSSIBLE HARM

- If decision is *punitive* in nature:
  - e.g. decide whom to deny bail based on risk of recidivism
  - Harm is caused when a protected group is given an unwarranted penalty
  - Heuristic: Use a fairness metric (separation) based on **false positive rate**
- If decision is *assistive* in nature:
  - e.g., decide who should receive a loan or a food subsidy
  - Harm is caused when a group in need is incorrectly denied assistance
  - Heuristic: Use a fairness metric based on **false negative rate**

# WHICH FAIRNESS CRITERIA?

- Decision: Classify whether a defendant should be detained
- Criteria: Anti-classification,

independence, or separation w/  
FPR or FNR?





# WHICH FAIRNESS CRITERIA?

**LOAN APPLICATION**

**Personal Information**

Name (Last) PUBLIC (First) JON (Middle Initial) J  
 Address (Mailing Address) 12345 MAIN STREET (City) ANCHORAGE (State) AK (Zip) 99501  
 E-Mail Address JQPJQPJQP@JQPJQP.COM (Home Telephone) 1111 1111 (Other Telephone) 222 222 2222

**Services needed**

UNDER REVIEW

**Current Income**

High School Graduate Or General Education (GED) Test Passed? Yes No  
 Highest grade completed  
 Military (Most recent first) Credits Earned  
 Quarterly or Other (Specify) Graduate Major or Subject

- Decision: Classify whether an applicant should be granted a loan.
- Criteria: Anti-classification, independence, or separation w/ FPR or FNR?

# WHICH FAIRNESS CRITERIA?

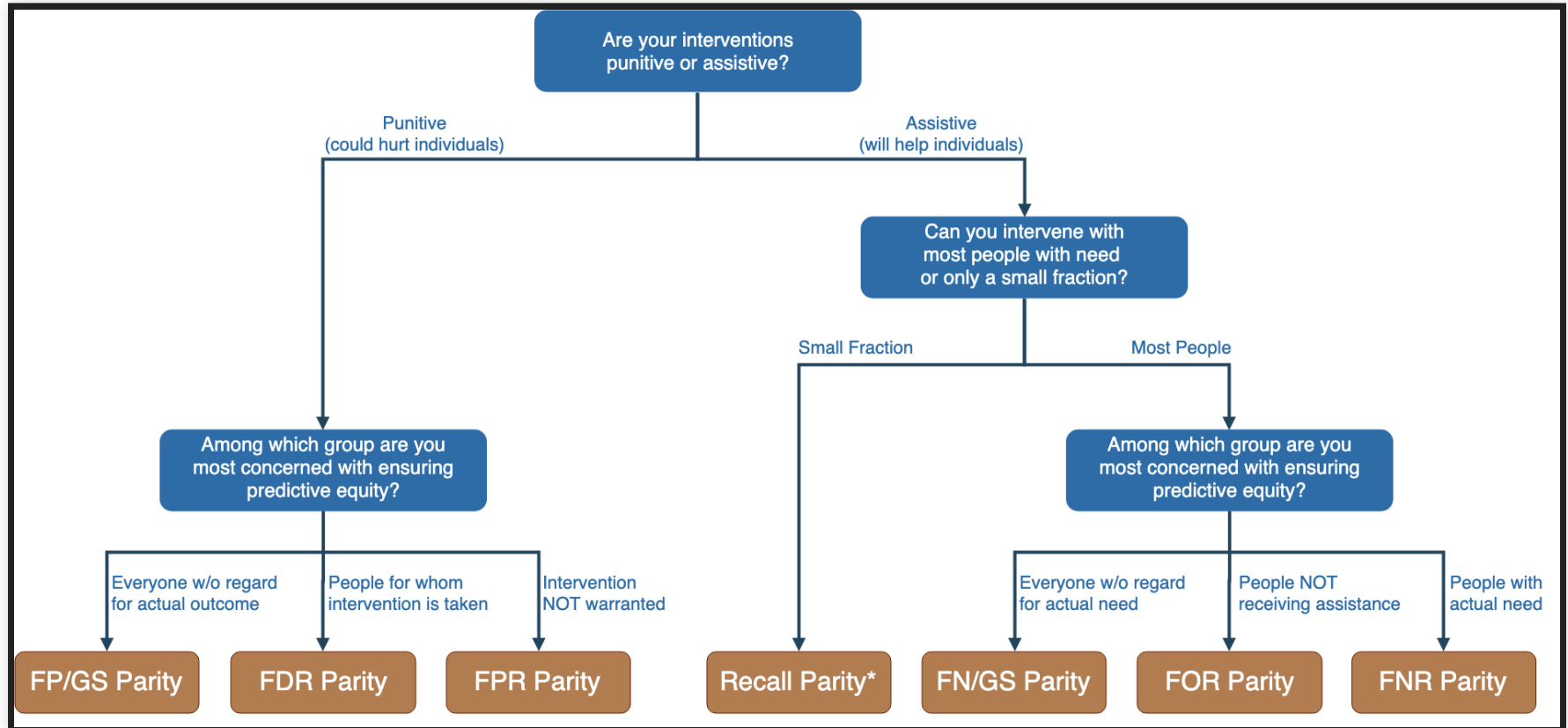


- Decision: Classify whether a patient has a high risk of cancer
- Criteria: Anti-classification, independence, or separation w/ FPR or FNR?





# FAIRNESS TREE



For details on other types of fairness metrics, see:  
<https://textbook.coleridgeinitiative.org/chap-bias.html>

# SUMMARY

- Definitions of fairness
  - Anti-classification, independence, separation
- Achieving fairness
  - Trade-offs between accuracy & fairness
- Achieving fairness as an activity throughout the entire development cycle
- Requirements engineering for fair ML systems
  - Stakeholders, sub-populations & unfair (dis-)advantages
  - Types of harms
  - Legal requirements

