ANITA B.ORG 2022 GRACE HOPPER CELEBRATION | NEXT IS NOW

SPEAKER

Identifying and Mitigating Bias in Machine Learning Models

Jana Gunn, M.S.

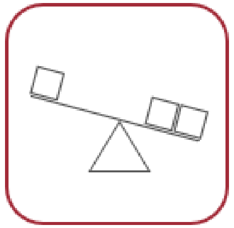
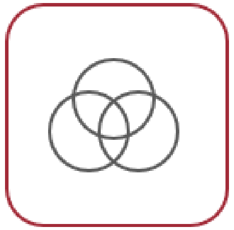
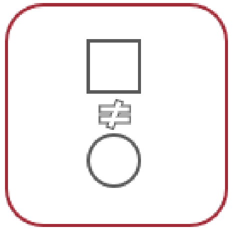
Sr. Director, Optum

Objectives

1. Understand why identifying and mitigating bias is important
2. Identify and discuss different definitions of fairness
3. Quantify harms found in machine learning models
4. Review strategies for mitigating bias in machine learning models

Any decision-making system can exhibit bias towards certain factors and thus, needs to be evaluated for fairness.

Sources of bias in machine learning

Data collection	Representation	Fit			Post processing
					
Historical	Representation	Measurement	Aggregation	Evaluation	Deployment
What may continue even with perfect sampling and feature selection	When population is under-represented and fails to generalize for sub-groups	When features make models blind and/or add factors that skew predictions	When subgroups are inappropriately combined	When there is unequal results representation or there are faulty performance metrics	When there is misuse or misinterpretation of predictive models
(E.g., racism)	(E.g., hiring predictions)	(E.g., proxy features)	(E.g., sex and race)	(E.g., A1C range variations by subclass; working history as indication of credit worthiness)	(E.g., use of diabetes risk score to determine donor eligibility; using a recidivism prediction model to determine length of sentence)

What is fairness?

Is the goal to achieve statistical parity or preserve preference?

Statistical Parity: Representative division regardless of preference

Preference Preservation: Strive to optimize for the preference of users

What is fairness?

Is the goal to avoid disparate treatment or disparate outcomes?

Procedural Fairness (equal opportunity): All individuals with similar characteristics should receive similar decisions from the system

Distributive Justice (equal outcome): All groups of individuals (as defined by specific identifiable characteristics such as race or sex) should receive similar outcomes from the system

Approaches to fairness

Fairness through Unawareness

Protected attributes are not used in prediction process

Pros

Guaranteed not to make an explicit judgment based on a sensitive attribute

Cons

Sometimes the attribute is important to the decision-making process (different symptoms based on the sex of a patient)

Unfairness (such as racism) can and does happen in "color-blind settings", and unawareness can mask and hide this.

It is entirely possible for an algorithm that has zero knowledge of the protected characteristic to be unfair and discriminatory.

Proxy variables are closely related to sensitive feature

Examples: hair length and gender, race and zip code

An algorithm trained on a dataset that embeds systemic bias will learn and perpetuate its patterns even if blind to the protected class.

Example: using healthcare costs as a proxy for disease severity

Individual Fairness

Similar outputs for similar individuals

Pros

Consistent between individuals

Cons

"Similarity" can be difficult to define, especially when multiple overlapping metrics are involved

There is no way to check whether group fairness is also satisfied under these definitions

Group Fairness

Predicts a particular outcome for individuals across groups with similar probability (e.g., a healthcare algorithm that assigns 5% of white patients as eligible for dialysis also assigns 5% of Black patients as eligible for dialysis)

Pros

Satisfies notions of avoiding penalizing or harming a specific group

Aligns with concerns about group equity (e.g., similar dialysis spending is granted to both Black and white patients)

Cons

No requirement to pick the "most qualified" within each group

Can be less accurate and potentially inappropriate if base rates of a label differ

Equality of Opportunity

Probability of an outcome is the same across different classes (e.g., if a man has a 40% chance of being hired for a job, so does a woman with similar experience)

Pros

True positive rate is the same for all groups

Cons

If base rates of the labels are different, there would be different false positive rates (e.g., if a higher proportion of women are qualified for the job, more unqualified men may be hired)

Counterfactual Fairness

Decisions for a person who is a member of group X are the same as they would be if that person were a member of group Y (e.g., the algorithm makes the same decision for a Black woman as it would have if she were a white woman)

Pros

Aligns with an intuitive and aspirational sense of fairness without being colorblind

Cons

Different factors are interrelated, and the world is too complex to build models that truly estimate the counterfactual

Defining the "similarly situated" member of the non-minority group can be difficult

Intersectional identities further the complexity

Fairness considerations

Data

Determining what data to use to train the model is a critical component when building a fair machine learning model.

Several factors should be considered

- Sample sizes of different groups (i.e., race, ethnicities)
- Representativeness of population
- Appropriateness of choice of label
- Label imbalance
- Adequate features for prediction
- Privacy considerations

Fairness considerations

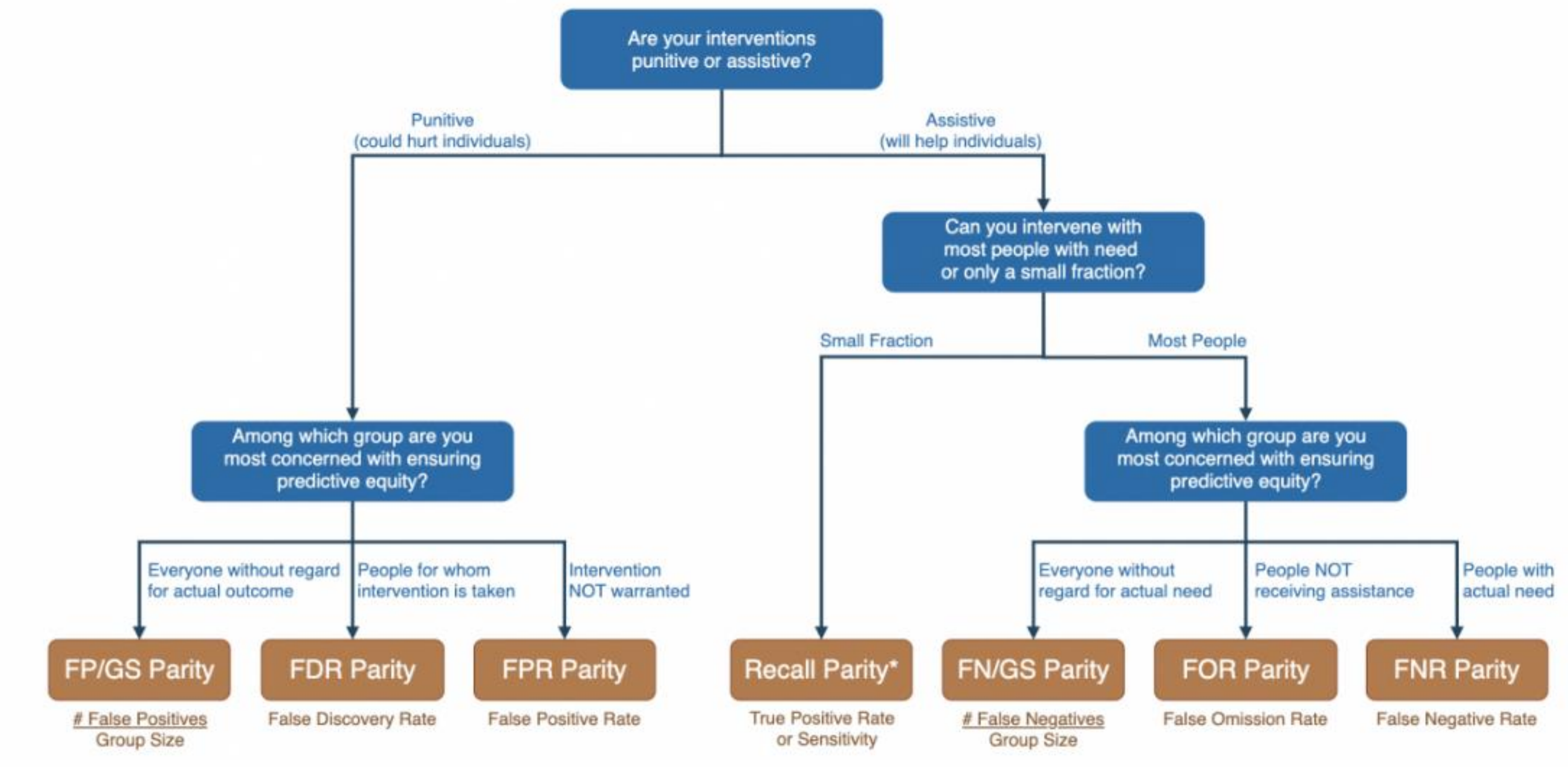
Types of harms

It is important to understand the types of harms the machine learning model could produce in production.

AI systems can exhibit one or more types of harm

- Allocation harms: Opportunities are unintentionally withheld from certain members of society
- Quality of service harms: System works well for some groups but not others
- Representation harms: System produces results that exacerbate stereotypes or under- represent certain populations

Fairness metrics



[Aequitas – Data Science and Public Policy (datasciencepublicpolicy.org)](datasciencepublicpolicy.org)

ANITA B.ORG 2022 GRACE HOPPER CELEBRATION SPEAKER

Mitigation Strategies

Within the model training stage, mitigation may occur at different steps.

1

Pre-Processing

A mitigation algorithm is applied to transform the input data to the training algorithm; for example, some strategies seek to remove the dependence between the input features and sensitive features

2

In-Processing

The model is trained by an optimization algorithm that seeks to satisfy fairness constraints.

3

Post-Processing

The output of a trained model is transformed to mitigate fairness issues; for example, the predicted probability of readmission is thresholded according to a group-specific threshold.

Pre-Processing

Removes the information correlated to the sensitive attributes while preserving as much information as possible

Pros

Preprocessed data can be used for any downstream task

No need to modify classifier

No need to access sensitive attributes at test time

Cons

Can only optimize certain metrics because it does not have the information of label Y

Inferior to the other two methods in terms of performance on accuracy and fairness measure

In-Processing

Add a constraint or a regularization term to the existing optimization objective

Pros

Good performance on accuracy and fairness measures

More flexibility to choose the trade-off between accuracy and fairness measures

No need to access sensitive attributes at test time

Cons

Need to modify classifier, which may not be possible in many scenarios

Post-Processing

Attempts to edit posteriors in a way that satisfies fairness constraints

Pros	Cons
Can be applied after any classifiers	Require test-time access to the protected attribute
Relatively good performance especially fairness measures	Lack the flexibility of picking any accuracy–fairness tradeoff
No need to modify classifier	

Recommended Tools

Tool	Functionality (Bias Detection)	Delivery (Visual Richness)	Usability (Ease of Use)	Mitigation (Integrated mitigation algorithms)
Fairlearn	✓	✓	✓	✓
Aequitas	✓	✓	✓	✗
IBM AI Fairness 360	✓	✓	✓	✓
Tensorflow	✓	✓	✓	✓

Summary

1. Machine learning models should demonstrate parity across identified sensitive groups.
2. It is important to not only identify disparities, but also mitigate them.
3. Mitigation can occur during pre-processing, at training time, or during post-processing.
4. Determining which algorithm to use is a tradeoff between performance and level of parity.

ANITA B.ORG 2022
GRACE HOPPER
CELEBRATION
SPEAKER

NEXT IS NOW

Thank you!

ANITA B.ORG 2022
GRACE HOPPER
CELEBRATION
SPEAKER

Appendix

Fairness Metrics

False Positive	FP_g	The number of entities of the group with $\hat{Y} = 1$ and $Y = 0$.
False Negative	FN_g	The number of entities of the group with $\hat{Y} = 0$ and $Y = 1$.
True Positive	TP_g	The number of entities of the group with $\hat{Y} = 1$ and $Y = 1$.
True Negative	TN_g	The number of entities of the group with $\hat{Y} = 0$ and $Y = 0$.
False Discovery Rate	$FDR_g = \frac{FP_g}{PP_g} = \Pr(Y = 0 \mid \hat{Y} = 1, A = a_i)$	The fraction of false positives of a group within the predicted positive of the group.
False Omission Rate	$FOR_g = \frac{FN_g}{PN_g} = \Pr(Y = 1 \mid \hat{Y} = 0, A = a_i)$	The fraction of false negatives of a group within the predicted negative of the group.
False Positive Rate	$FPR_g = \frac{FP_g}{LN_g} = \Pr(\hat{Y} = 1 \mid Y = 0, A = a_i)$	The fraction of false positives of a group within the labeled negative of the group.
False Negative Rate	$FNR_g = \frac{FN_g}{LP_g} = \Pr(\hat{Y} = 0 \mid Y = 1, A = a_i)$	The fraction of false negatives of a group within the labeled positives of the group.

Fairlearn
[Fairlearn]

Aequitas
[GitHub - dssg/aequitas: Bias and Fairness Audit Toolkit]

IBM AI Fairness 360
[AI Fairness 360 (mybluemix.net)]

Tensorflow
[Responsible AI Toolkit | TensorFlow]