AI Fairness and Biases

Jin L.C. Guo
SOCS, McGill University

What is the harm

Quality of Service: degraded user experience

Harm of allocation: withhold opportunity or resources

Harm of representation: reinforce subordination along the line of identity, stereotype

What kind of harm your system might cause? To whom?

Legally Recognized Protected Classes

United States federal anti-discrimination law:

- Race – Civil Rights Act of 1964
- Religion – Civil Rights Act of 1964
- National origin – Civil Rights Act of 1964
- Age (40 and over) – Age Discrimination in Employment Act of 1967
- Sex – Equal Pay Act of 1963 and Civil Rights Act of 1964
  - Sexual orientation and gender identity as of Bostock v. Clayton
- County – Civil Rights Act of 1964
- Pregnancy – Pregnancy Discrimination Act

Legally Recognized Protected Classes

- Familial status – Civil Rights Act of 1968 Title VIII: Prohibits discrimination for having children, with an exception for senior housing. Also prohibits making a preference for those with children.
- Disability status – Rehabilitation Act of 1973 and Americans with Disabilities Act of 1990
- Veteran status – Vietnam Era Veterans' Readjustment Assistance Act of 1974 and Uniformed Services Employment and Reemployment Rights Act
- Genetic information – Genetic Information Nondiscrimination Act

More than legally protected classes

Other societal categories like location, topical interests, (sub)culture etc.

Subpopulations may be application-specific, intersectional, subject to complex social constructs

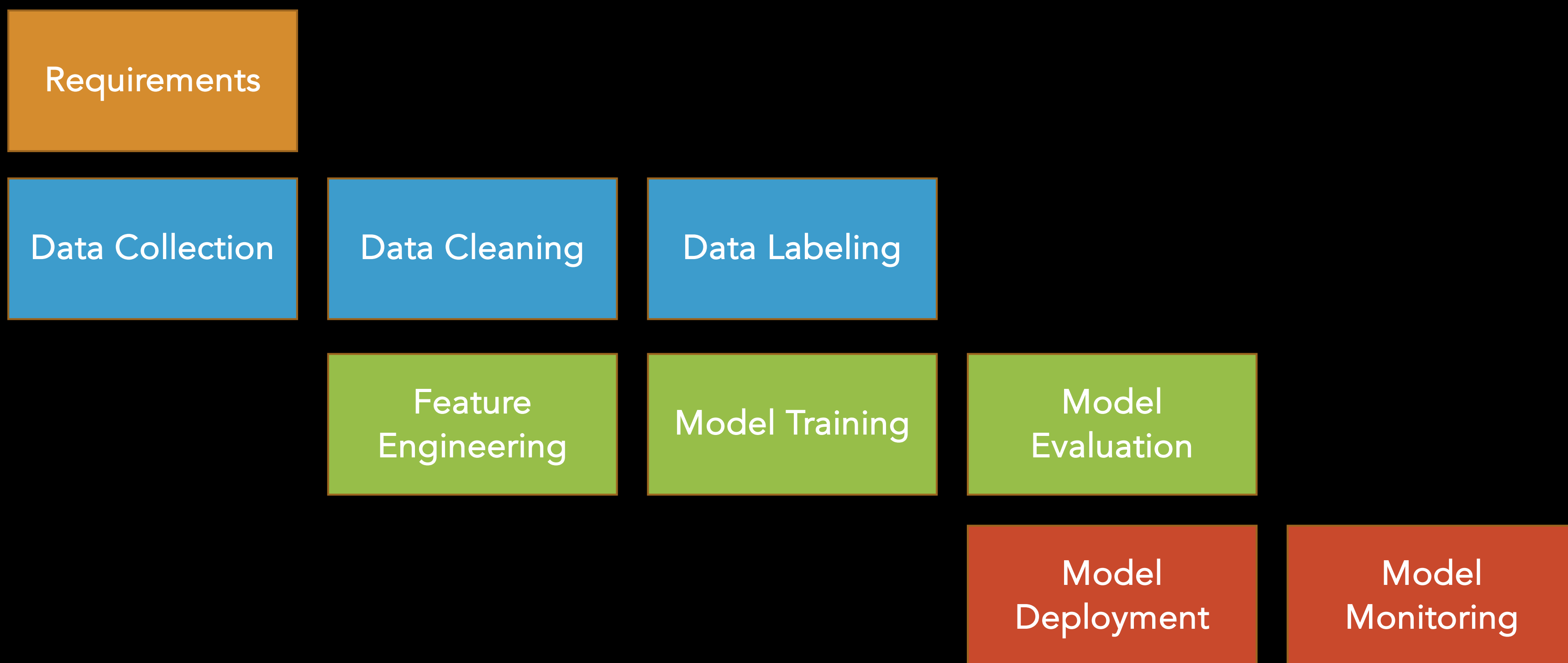
> *"Most of the time, people start thinking about attributes like [ethnicity and gender...]. But the biggest problem I found is that these cohorts should be defined based on the domain and problem. For example, for [automated writing evaluation] maybe it should be defined based on [...whether the person is] a native speaker."*

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.

Activity: Case study for a ML based hiring system

- Pick up a hiring system for a concrete domain (educational institution, tech company, government, etc.). Consider what is the goal of the hiring system for human first?
- Then consider an ML based such hiring system
    - Where are the sources of biases?
    - What are the potential harms for different stakeholder groups when they are treated with biases?
    - How do you plan to mitigate them?

Sources of Biases and Mitigation Strategies for ML Pipeline



Sources of Biases and Mitigation Strate

Requirer

Data Collection

Data Cleaning

Data Labeling

1.1.b Scrutinize resulting system vision for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)
- Tradeoffs between expected benefits and potential harms for different stakeholder groups
  - Consider who the system will give power to and who it will take power from
  - Consider which expected benefits you are willing to sacrifice to mitigate potential harms

1.2.a Solicit input on system vision and potential fairness-related harms from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
  - Consider whether any stakeholder groups would prefer that the system not exist or not be deployed in all contexts, what alternatives they would prefer, and why
- Domain or subject-matter experts
- Team members and other employees

Model Deployment

Model Monitoring

Sources of Biases and Mitigation Strate

Requirem

Data Colle

2.2.a Define datasets needed to develop and test the system, considering:

- Desired quantities and characteristics, considering:
  - Relevant stakeholder groups, including demographic groups
    - Consider oversampling smaller stakeholder groups, but be aware of overburdening
  - Expected deployment contexts
- Potential sources of data
  - Consider reviewing all datasets from third-party vendors
- Collection, aggregation, or curation processes, including:
  - Procedures for obtaining meaningful consent from data subjects
  - People involved in collection, aggregation, or curation, including demographic groups
    - Consider whether people involved might introduce societal biases
  - Incentives for data subjects and people involved in collection, aggregation, or curation
    - Consider whether data subjects might feel undue pressure to provide data
- Software, hardware, or infrastructure involved in collection, aggregation, or curation
- … … (Regulations, Assumptions)

Sources of Biases and Mitigation Strategies for ML Pipeline

Requirem

Data Col

Feature

Model

2.3. Based on potential fairness-related harms identified so far, define fairness criteria, considering:

- How criteria will be assessed (e.g., fairness metrics and benchmark dataset, system walkthroughs with diverse stakeholders or personas) at each subsequent stage of the lifecycle, including
    - People involved in assessment (e.g., judges), including demographic groups
    - Datasets needed to assess fairness criteria
- Acceptable (levels of) deviation from fairness criteria
- Potential adversarial threats or attacks to fairness criteria
- Assumptions made when operationalizing system vision via fairness criteria
    - Consider whether these assumptions are sufficiently well justified

3.3.a Assess fairness criteria according to fairness criteria definitions, considering:

- Acceptable (levels of) deviation from fairness criteria
- Tradeoffs between different fairness criteria
- Tradeoffs between performance metrics and fairness criteria
- Discrepancies between development environment and expected deployment contexts

Sourc
Strate

Requirer

Data Coll

5.1 Participate in public benchmarks

5.1.a Participate in public benchmarks so that stakeholders can contextualize system performance

5.1.b Revise system to mitigate any harms revealed by benchmarks; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting deployment

5.2 Enable functionality for stakeholder feedback

5.2.a Establish processes for responding to or escalating stakeholder feedback, including:

- Stakeholder comments or concerns
- Third-party audits

6.1 Monitor deployment contexts

6.1.a Monitor deployment contexts for deviation from expectations, including:

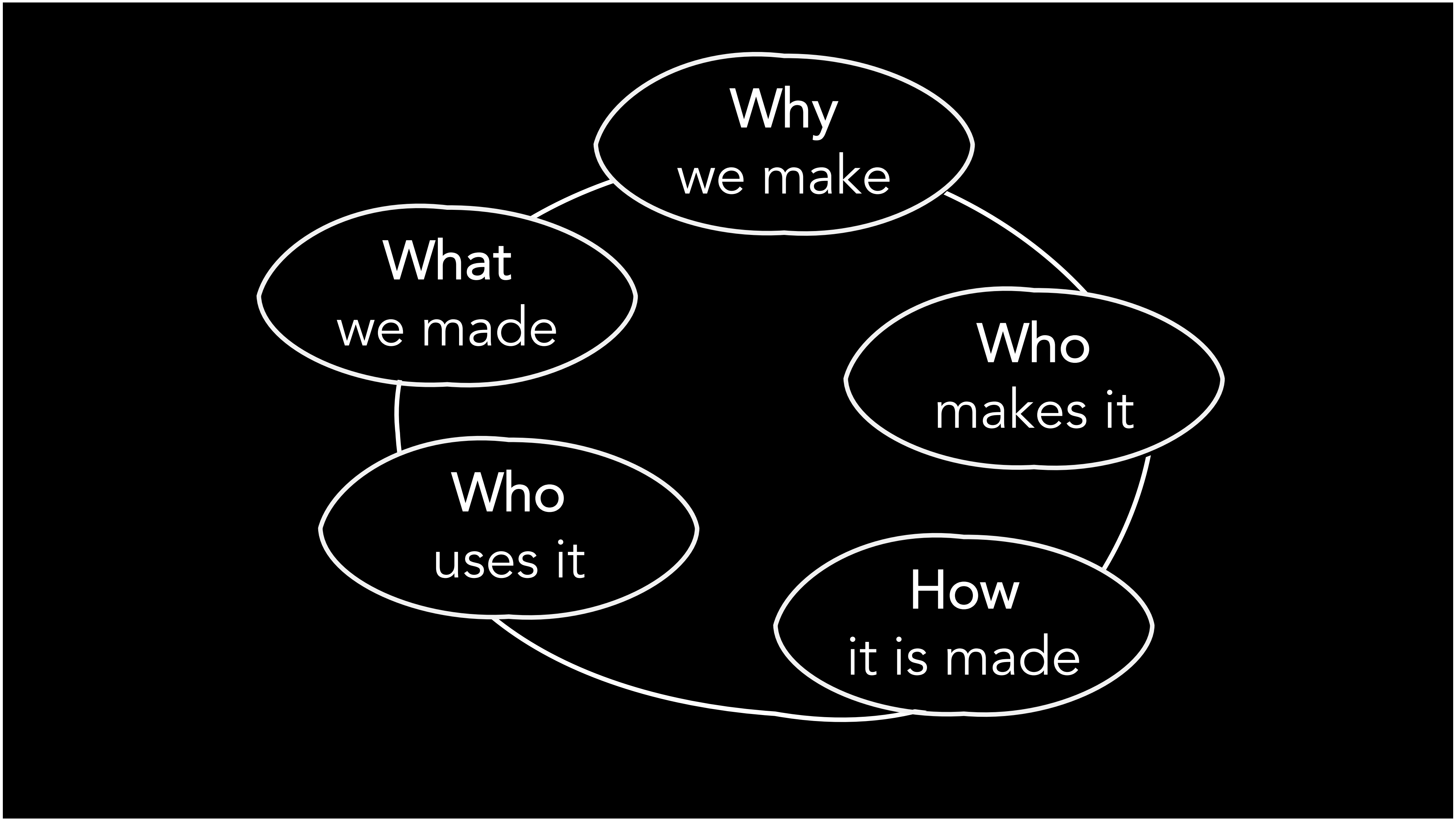
- Unanticipated stakeholder groups, including demographic groups
- Adversarial threats or attacks

6.1.b Revise system (including datasets) to match actual deployment contexts; if this is not possible, document why, along with expected impacts on stakeholders, and consider rollback or shutdown

6.2 Monitor fairness criteria

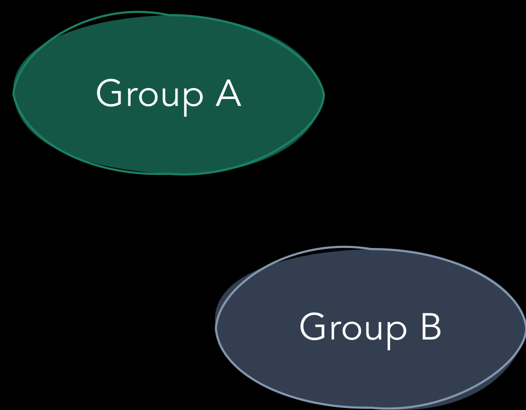
6.3 Monitor stakeholder feedback

6.4 Revise system at regular intervals to capture changes in societal norms and expectations


Why
we make
What
we made
Who
makes it
Who
uses it
How
it is made

Measuring Fairness

- Group Fairness – based on statistical parity



Error Rate (false positive, false negative)

Favored Outcome

Measuring Fairness

- Individual Fairness

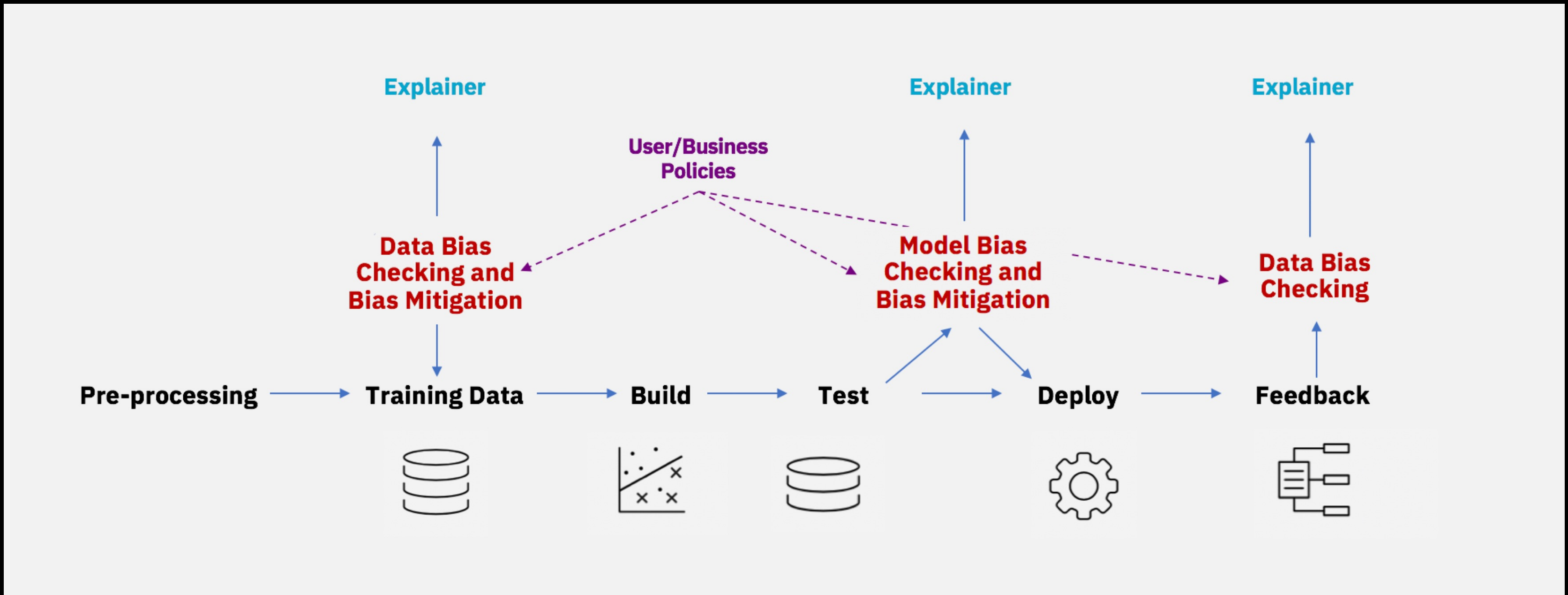
Individual *i*

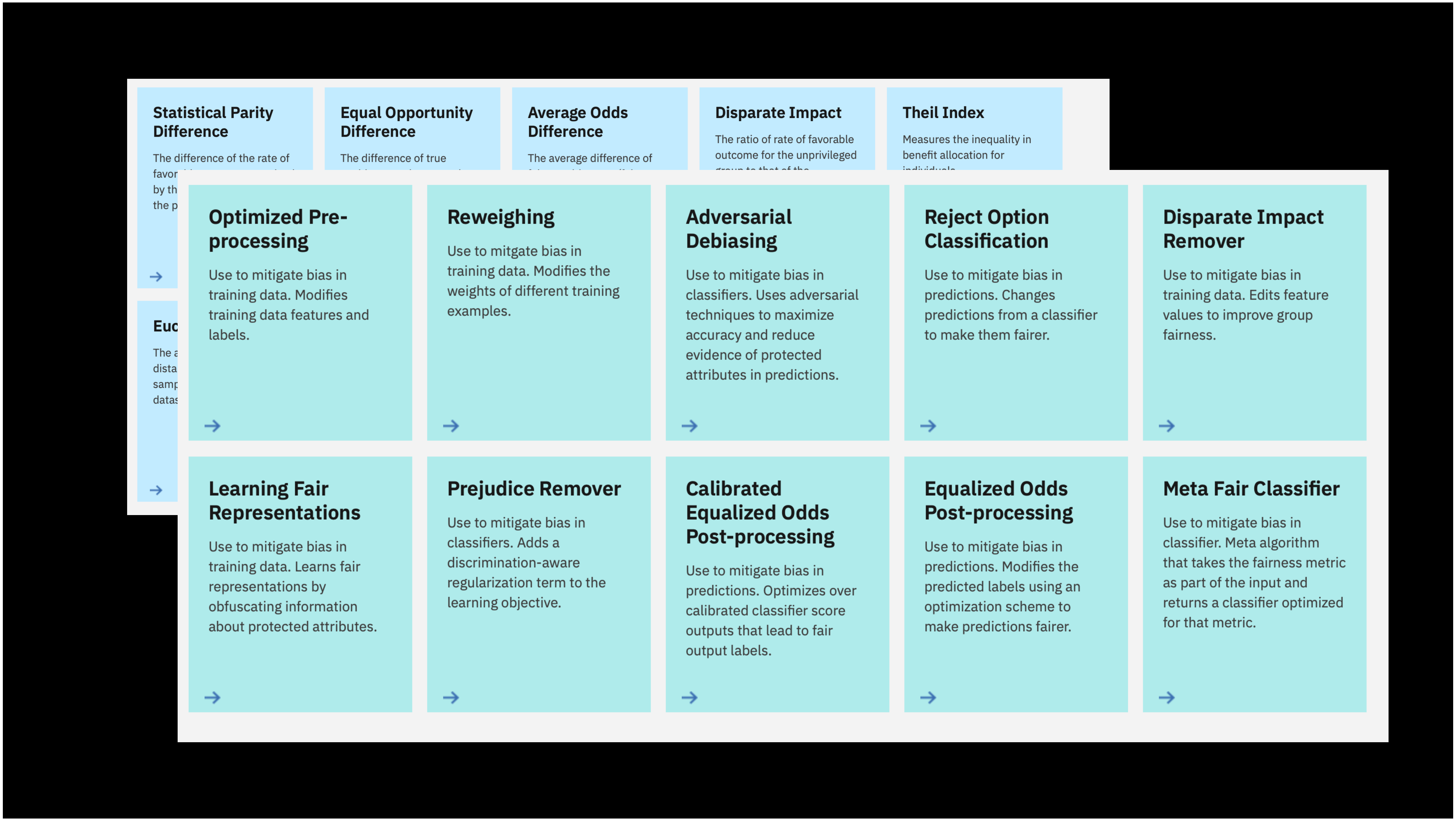
Individual *j*

Individuals who are similar (with respect to the task)
Should be treated similarly.

Fairness Toolkits

AI Fairness 360





Statistical Parity Difference
The difference of the rate of
Equal Opportunity Difference
The difference of true
Average Odds Difference
The average difference of
Disparate Impact
The ratio of rate of favorable outcome for the unprivileged
Theil Index
Measures the inequality in benefit allocation for
Optimized Pre-processing
Use to mitigate bias in training data. Modifies training data features and labels.
Reweighing
Use to mitigate bias in training data. Modifies the weights of different training examples.
Adversarial Debiasing
Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.
Reject Option Classification
Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.
Disparate Impact Remover
Use to mitigate bias in training data. Edits feature values to improve group fairness.
Learning Fair Representations
Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.
Prejudice Remover
Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.
Calibrated Equalized Odds Post-processing
Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.
Equalized Odds Post-processing
Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.
Meta Fair Classifier
Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

Fairlearn

algorithm	description	binary classification	regression	supported fairness definitions
`fairlearn.reductions.ExponentiatedGradient`	A wrapper (reduction) approach to fair classification described in *A Reductions Approach to Fair Classification* [5].	✔	✔	DP, EO, TPRP, FPRP, ERP, BGL
`fairlearn.reductions.GridSearch`	A wrapper (reduction) approach described in Section 3.4 of *A Reductions Approach to Fair Classification* [5]. For regression it acts as a grid-search variant of the algorithm described in Section 5 of *Fair Regression: Quantitative Definitions and Reduction-based Algorithms* [4].	✔	✔	DP, EO, TPRP, FPRP, ERP, BGL
`fairlearn.postprocessing.ThresholdOptimizer`	Postprocessing algorithm based on the paper *Equality of Opportunity in Supervised Learning* [6]. This technique takes as input an existing classifier and the sensitive feature, and derives a monotone transformation of the classifier's prediction to enforce the specified parity constraints.	✔	✘	DP, EO, TPRP, FPRP

						Models covered				Group fairness						Individual			
Tool	Setup	Open source user license	Release date	Organization	Open for anyone to contribute code?	Regression	Classification (binaryoutcome)	Multi-class outcome	Handles multi-class protected feature?	Demographic parity (statistical parity)	Equal opportunity / True positive parity / false positive error rate balance	Equal odds (True positive and false positive parity)	Disparate impact	Discovery rate	Omission rate	Counterfactual fairness	Sample distortion metrics	Other fairness metrics	Bias mitigation
Scikit-fairness / scikit-lego	python (sklearn)	MIT	2019-03-31	N/A	✓	✓	✓	X	X	✓	✓	X	X	X	X	X	X	N/A	Pre-processing: information filter
IBM Fairness 360	python 3.5+, R	Apache 2.0	2018-06-01	IBM	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	Generalized Entropy Index Differential Fairness and Bias Amplification (full list here: https://aif360.readthedocs.io/en/latest/modules/generated/aif360.metrics.ClassificationMetric.html)	Optimized Preprocessing, Disparate Impact Remover, Equalized Odds Post-processing, Reweighing, Reject Option Classification, Prejudice Remover Regularizer, Calibrated Equalized Odds Postprocessing, Learning Fair Representations, Adversarial Debiasing, Meta-Algorithm for Fair Classification, Rich Subgroup Fairness
Aequitas tool	python 3.6+	Custom	2018-02-13	UChicago	✓	X	✓	X	✓	✓	✓	✓	X	✓	✓	X	X	N/A	N/A
Google What-if tool	Tensorboard / Jupyter or Colab notebook	Apache 2.0	2018-09-11	Google	✓	✓	✓	✓	✓	✓	✓	X	X	X	X	✓	X	Group thresholds	Threshold optimization based on fairness constraints
PyMetrics audit-ai	python	MIT	2018-05-18	PyMetrics	X	✓	✓	X	X	X	X	X	✓	X	X	X	X	Statistical tests to determine chance the disparity is due to random chance (ANOVA, 4/5th, fisher, z-test, bayes factor, chi squared sim beta ratio, classifier posterior_probabilities)	N/A
Fairlearn	python	MIT	2018-05-15	Microsoft	✓	✓	✓	X	✓	✓	✓	✓	X	X	X	X	X	Group max / min / summary	Exponentiated Gradient, GridSearch, Threshold Optimizer

Figure 1: Open source toolkit feature summary table

Lee, M.S.A. and Singh, J., 2021, May. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Key Takeaways

- Fairness is tightly connected to other principles such as auditability, privacy
- Fairness is relevant to every stage of the ML pipeline, starting from the scoping to monitoring
- Consider and involve diverse stakeholders at various stages

Next

Design for Creativity