Fairness and Biases

Jin Guo Nov 24th 2020

What is the harm?

Harm of allocation: withhold opportunity or resources

Quality of Service: degraded user experience

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
- 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

https://en.wikipedia.org/wiki/Protected_group

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.

Sources of Biases and Mitigation Strategies

Requirements

Data Collection

Data Cleaning

Data Labeling

Feature Engineering

Model Training

Model Evaluation

Model Deployment

Model Monitoring

Sources of Biases and Mitigation Strategies

Requiremen

Data Collect

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

Madaio, Michael A., Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-14. 2020.

Sources of Biases and Mitigation Strategies

Requiremen

Data Collect

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)

Madaio, Michael A., Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach.
"Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI."
In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-14. 2020.

Sources of Biases and Mitigation Strategies

Requireme

Data Collec

2.3.a 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

Madaio, Michael A., Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach.
"Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI."
In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-14. 2020.

Sourc

Requiren

Data Collection

Data Cleaning

Data Labeling

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

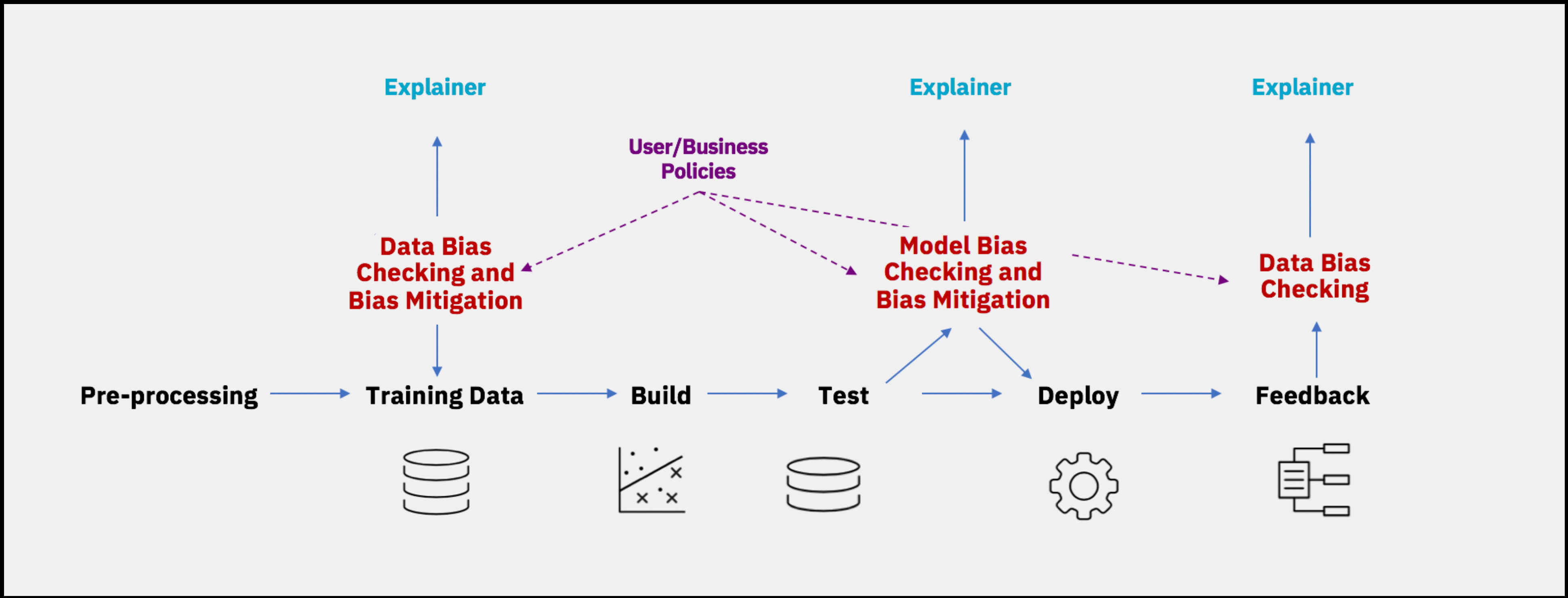
6.3 Monitor stakeholder feedback

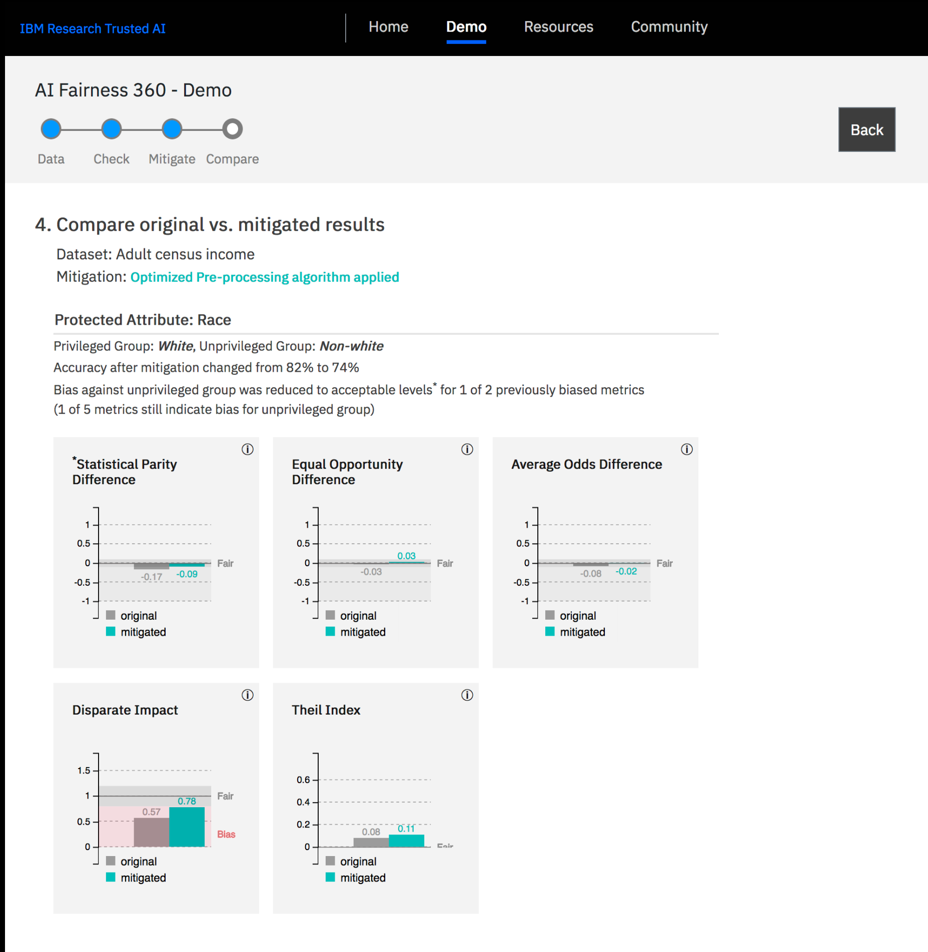
6.4 Revise system at regular intervals to capture changes in societal norms and expectations

Madaio, Michael A., Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach.
"Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI."
In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-14. 2020.

Tools

AI Fairness 360





Supported bias mitigation algorithms

- Optimized Preprocessing (Calmon et al., 2017)
- Disparate Impact Remover (Feldman et al., 2015)
- Equalized Odds Postprocessing (Hardt et al., 2016)
- Reweighing (Kamiran and Calders, 2012)
- Reject Option Classification (Kamiran et al., 2012)
- Prejudice Remover Regularizer (Kamishima et al., 2012)
- Calibrated Equalized Odds Postprocessing (Pleiss et al., 2017)
- Learning Fair Representations (Zemel et al., 2013)
- Adversarial Debiasing (Zhang et al., 2018)
- Meta-Algorithm for Fair Classification (Celis et al.. 2018)
- Rich Subgroup Fairness (Kearns, Neel, Roth, Wu, 2018)

https://github.com/Trusted-AI/AIF360

Tools

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

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