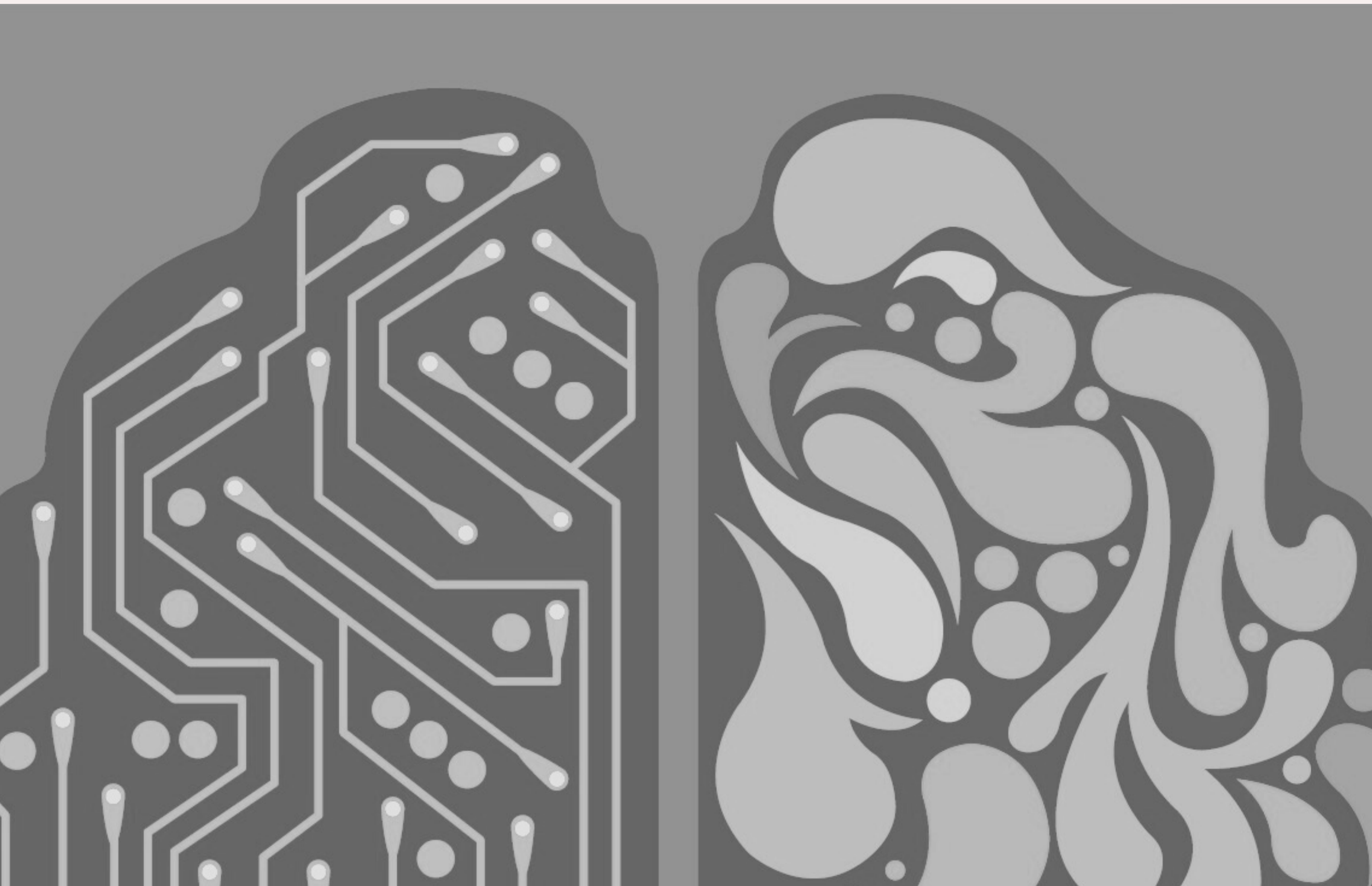
DEBIASING YOUR PRODUCT

RIANA SHAH



SUPPORTED BY

WE ARE SO GRATEFUL TO ALL THE INSTITUTIONS THAT HAVE PROVIDED SUPPORT TO US ALONG THE WAY!

Massachusetts
Institute of
Technology

HARVARD
Kennedy School

we
WE ARE FAMILY
FOUNDATION

Hi
Harvard
innovation lab

MIT
sandbox
Innovation Fund Program

Edmund
Hillary
Fellowship

Stanford
University

SWARTHMORE

TABLE OF CONTENTS

INTRODUCTION

Hello there,

Welcome to the EthixAI guide to debiasing your product! EthixAI is an initiative based at MIT and Harvard that combats algorithmic bias to fight systemic bias in technology.

During my graduate school career I served as MIT's Senate head of Diversity, Equity and Inclusion and ran the AI Ethics Reading Group. Throughout my time in those two roles, I struggled to find an accessible guide to debiasing machine learning products.

The goal of this e-book is to provide readers with a general guide to bias in artificial intelligence algorithms. You do not need to be a coder or a sociologist to gain value from this and I hope it is valuable to both groups of individuals.

This guide is divided into three sections:
1. Understanding Bias, which defines the different kinds of bias and how they arise,
2. Accounting for Bias, which provides systematic methods of identifying bias during product development
3. Mitigating Bias, which includes an Algorithmic Audit and gives strategies to actively combat bias in our AI algorithms and products.

We also have some case studies in the appendix and further reading. Use this guide as a kicking off point to debias your product!

In solidarity,
Riana Shah

Roadmap

This guide is divided into three sections:

1

UNDERSTANDING BIAS

Defining the many kinds of bias and how they arise, as well as the problems they may bring

2

ACCOUNTING FOR BIAS

Systematic methods of identifying bias during product development

3

MITIGATING BIAS: ALGORITHMIC AUDIT

Strategies to actively combat the bias in our AI algorithms and products

ETHIXAI TEAM

WE ARE SO GRATEFUL TO ALL THE INDIVIDUALS WHO WORKED HARD TO MAKE THIS PROJECT A REALITY AND THE ADVISORS WHO GAVE US CRUCIAL FEEDBACK ALONG THE WAY.

Riana Shah

Christie Little

Samuel Rothstein

Eishna Rangnathan

Maria Fernanda Sampaio Ferreira

Natasha Markov-Riss

Devyani Mahajan

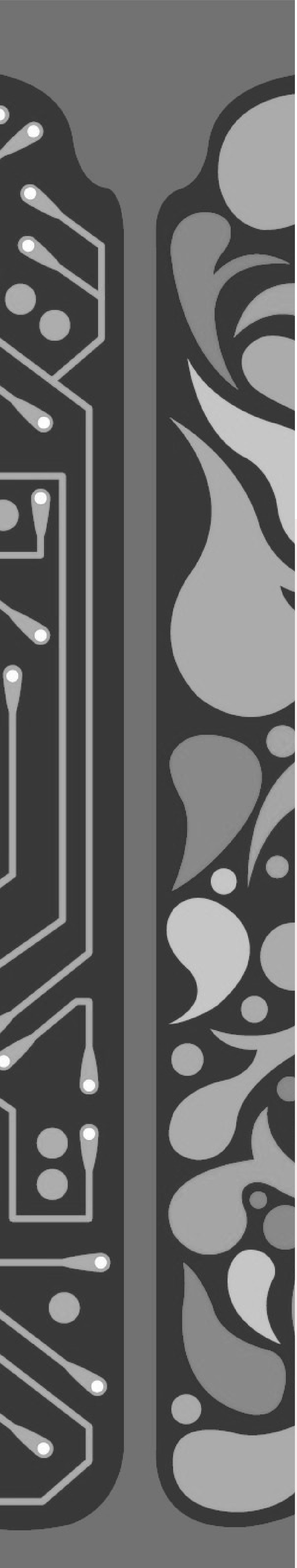
Advisors

Matthew Rhodes Kropf, MIT

Kathy Pham, Harvard & The White House

Ben Mitchell, Swarthmore

John Akula, MIT



PART 1: UNDERSTANDING BIAS IN AI

"The rise of machine learning is every bit as far-reaching as the rise of computing itself... a vast new infrastructure of techniques are emerging and we are just learning their full capabilities "

-Kate Crawford, Researcher

AI-based systems are used to make decisions that have a tremendous impact on individuals and society as a whole. It is critical to move beyond traditional AI algorithms that are optimized solely for predictive performance. Algorithms should strive to embed ethical and legal principles in their design, training and implementation. AI has the potential to propagate pre-existing biases and evolve criteria for new types of biases.

To build a debiased tech-product, we must **understand bias**, **mitigate bias**, and **account for bias:**

- To **understand bias**, we must derive approaches to determine how bias is conceived in society and use this sociological understanding to train our AI algorithms

- To **identify or account for bias**, we must formulate methods in three distinct stages of AI decision making:
    1. Data inputs
    2. Learning algorithms
    3. Model output representation.

- To **mitigate bias**, we must implement strategies for combating possible biases we've identified and cultivate an environment with increased representation and diversity to correct for these biases from the start.

OUR WORKING DEFINITIONS OF BIAS

1 IMPLICIT BIAS

A disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair. Implicit biases can be innate or learned. People may develop biases for or against an individual, a group, or a belief.

2 SYSTEMIC BIAS

The inherent tendency of a structure to favor certain bodies or outcomes, as created and reinforced by social, governmental, political, and economic systems in place.

3 STATISTICAL BIAS

This can involve either bias in data collection or in the difference between the expected value and actual value of an estimate. With respect to the former, this is when the sample isn't representative of the population of interest. The latter is with respect to "results that are systematically off the mark."

OUR WORKING DEFINITION OF BIAS

4 MACHINE LEARNING BIAS

Machine learning bias refers to when a machine learning model's erroneous assumptions results in systemically biased predictions.

It reflects and amplifies implicit biases, and reflects and fuels the structures that allow systemic biases

a INDUCTIVE BIAS

In machine learning, inductive bias refers to the idea that an algorithm must make assumptions and therefore have biases in order to generalize from training data to novel data. This is because for any training dataset, there are an infinite number of functions to choose from that would fit this dataset. Therefore, in order to select which function to use to generalise from training to novel data, assumptions and therefore biases must come into play.

b UNDERFITTING

Underfitting in machine learning is related to a property called the bias-variance tradeoff. In this context, bias refers to when there are erroneous assumptions being made by the model which lead the model to miss important relations between the features and target labels. This bias error is known as underfitting.

WHY YOU SHOULD CARE ABOUT THIS

LEGAL CASE

- Federal Law prohibits discrimination based on race, sex, national origin, religion, age, or disability are protected classes. If an algorithm discriminates against any of these protected classes, the makers of the algorithm may be held liable.

- The state of California just passed CPRA, which ups the ante on privacy restrictions. CPRA also has provisions about decision-making models for algorithms, stating that decision-making models for algorithms must be explainable.

  - Under CPRA, California consumers have the right to be excluded from algorithmic-based decision making. Companies must be able to facilitate non-algorithmic decision making for all processes.

  - While CPRA legislation is specific to California, it is clear that the direction of regulations is moving towards forcing explainability in AI algorithms.

- The GDPR requires that decisions made by an algorithm be explainable, and your algorithm is no exception.

WHY YOU SHOULD CARE ABOUT THIS

REPUTATIONAL CASE

- Offering users a tool or product that is based on a biased AI algorithm can result in reputational harm for your organisation. If users believe a product they are using is biased toward them because of a protected characteristic their trust toward the company can decrease.

- As an example, Amazon's AI recruiting tool stopped being used after it was found that the tool favoured men over women for technical roles. If women believe they may be discriminated against by a particular algorithm they may be less likely to apply to work at the firm.

BUILDING ROBUST PRODUCTS CASE

- As AI gets used more and more in services, de-biasing these models is now an essential component of providing clients with a robust product.
- Transparency with clients about the decision-making process and assumptions behind any machine learning model used in your services is a key pillar to build credibility with clients and grow your business.

HOW DOES BIAS ARISE?

Bias can become present in an algorithm long before data is collected as well as during other stages of the training and development process. Bias in AI can be a chicken and an egg problem - is it the algorithm or is it the training data or some combination of both?

First, when a computer scientist creates a machine learning model, their first task is to decide what problem they wish to optimize. If bias is not considered during this process, the machine learning model may use biased assumptions to optimize the predictive model.

Second, data collection plays a crucial role in creating a machine learning model. There are two ways that bias can show up in training data:

1. If the data collected is unrepresentative of reality, training data will likely be biased
2. If the training data reflects existing prejudices then the algorithm will also reflect these prejudices.

Third, preparation can also introduce bias into an AI algorithm. During data preparation, a computer scientist selects the attribute that they want the algorithm to consider. Depending on these choices, bias can be introduced into the AI algorithm at this stage of the process.

Hence, you need to watch out for bias at various stage in the algorithm development process.

PART 2: ACCOUNTING FOR BIAS

"Sometimes respecting people means making sure your systems are inclusive such as in the case of using AI for precision medicine, at times it means respecting people's privacy by not collecting any data, and it always means respecting the dignity of an individual."

-Joy Buolamwini, Computer Scientist

ACCOUNTING FOR BIAS

MANY APPROACHES FOR ACCOUNTING FOR BIAS FALL INTO ONE OF THREE CATEGORIES

A. PRE-PROCESSING METHODS

B. IN-PROCESSING METHODS

C. POST-PROCESSING METHODS

PRE-PROCESSING METHODS

FOCUSED ON THE DATASET

This section focuses on pre-processing methods to debias your machine learning algorithm, which are methods primarily focused on transforming the dataset. Following are a few of the guiding notions behind pre-processing methods.

- Protected attributes should not influence the output of a machine learning model.
- A less biased training dataset will result in a less biased algorithm.
- Balance the training dataset while minimizing data interventions.
    - This approach is portable, as a balanced dataset can be used to train any machine learning algorithm.

Example of Methods

1. **Counterfactual Fairness:** Modify the original data distribution by altering class labels.
2. **Reweighing**: Assign different weights to instances based on their features.
3. **Labelling**: Important to have group of people with diverse demographics labelling the dataset.
4. **Sampling**: Increase the representation of minority groups in the dataset by resampling the dataset.
5. **Remove protected attributes:** Remove all protected attributes as well as features they are highly correlated with from your feature space.

PRE-PROCESSING METHODS

	DESCRIPTION	EXAMPLE
COUNTER-FACTUAL FAIRNESS	Intuition • Ensures model outputs are the same in actual world and in a counterfactual world where individual belongs to a different demographic. • Helpful for us to understand why the model predicted certain labels for certain objects and whether the reasoning is discriminatory. How it works • If there is a protected attribute, ie. sex, influencing the results, change the labels of some objects with sex = female from – to +, and labels of same number of objects with sex = male from + to – • Keeps overall positive class ratio • Choose to relabel objects that were closest to the decision border	Use Case • Project by the Alan Turing Institute • Used counterfactual fairness to test whether the algorithm used by judges and parole officers in the US to score a defendants likelihood of reoffending is racially biased. • Found racial bias in the algorithm. To learn more • Kusner et. al., Counterfactual Fairness (2017) • Calders et. al., Building classifiers with independency constraints (2009)
REWEIGHING	How it works • Assign weights to tuples in the training dataset to remove discrimination. • If there is a protected attribute ie. sex that is influencing results, give objects with sex = female and + class a higher weight than objects with sex = female and – class. Similarly, give objects with sex = male and – class higher weights than objects with sex = male and + class. • Keeps overall positive class ratio • Reduces dependency with minimal number of changes to dataset	Use Case • Consider a sample of job applications. To remove the dependency between the Sex and Class labels (positive or negative), we can assign weights. • We have four possible combinations: male +, male -, female + and female -. So, we only need to assign four weights. • Then, we weight each individual application based on which of the four possible combinations it falls into. To learn more • Calders et. al., Building classifiers with independency constraints (2009)

PRE-PROCESSING METHODS

DESCRIPTION | EXAMPLE

LABELLING

Description

Intuition

- People have implicit biases and by including people from diverse backgrounds, reduces chance that these biases will affect dataset labels

How it works

- Ensure each object is labelled by multiple people with diverse demographics

Example

Use Case

- Casual Conversations is an open-source Facebook dataset of tens of thousands of videos of participants having conversations.
- Facebook hired a group of trained, ethnically diverse annotators from around the country to label participant's apparent skin tones and the ambient lighting conditions.
- These labels will later be used to determine how AI systems perform across different skin tones and low-light ambient conditions.

SAMPLING

Description

Intuition

- Use oversampling and undersampling to decrease sampling bias in your dataset.
- Especially useful for classifier learners that can't incorporate weights in their learning process

How it works

- Split your data into four classes based on a sensitive attribute: favoured group with positive label, favoured group with negative label, unfavoured group with positive label, unfavoured group with negative label
- Compute expected size of each of the four groups if dataset were non-discriminatory and under-sample some groups while oversampling others.

Example

Use Case

- Consider a sample of job applications with data on sex and the class label (+ or –) of applicants, among other features. Let female sex be the deprived group and male sex the favoured group. First we look at how many favoured +, favoured –, deprived + and deprived – applications we have. In a world without discrimination, these four groups would have an equal number of members. So, we can undersample and oversample accordingly to adjust for this.

To learn more

- Kamiran et. al., Data preprocessing techniques for classification without discrimination (2012)

PRE-PROCESSING METHODS

DESCRIPTION

EXAMPLE

REMOVE PROTECTED ATTRIBUTES

Intuition

- If you don't remove protected attributes from your feature space, these biased features will influence your model's predictions.
- Removing protected attributes alone is insufficient, as other features in the dataset may be highly correlated with these attributes.
  - This means that your model can still produce biased results by basing its predictions on these highly correlated features.

How it works

- Find features which are highly correlated with each protected attribute, and remove both the protected attribute and these highly correlated features.
- A limitation of this method is that sometimes you might need these correlated features, and therefore removing the protected attributes makes it difficult to correct for the attribute's bias in your predictions.

Use Case

- An example would be an algorithm used for loan default prediction with race and zip code as two of the features in the dataset. Race is a protected attribute and must therefore be removed. However, zip code is frequently highly correlated with race, so even if race is removed, racial biases can still be learned by the algorithm through the zip code. So, in this example it would also be necessary to remove each data instances zip code from the dataset.

ACCOUNTING FOR BIAS

MANY APPROACHES FOR ACCOUNTING FOR BIAS FALL INTO ONE OF THREE CATEGORIES

A. PRE-PROCESSING METHODS

B. IN-PROCESSING METHODS

C. POST-PROCESSING METHODS

IN-PROCESSING METHODS

FOCUSED ON THE MACHINE LEARNING ALGORITHM

This section focuses on in-processing methods to debias your machine learning algorithm, which are methods primarily focused on adjusting your machine learning algorithm. In-processing methods are inherently algorithm-specific. Therefore, research must be done on which debasing methods are best suited for the specific algorithm used in your project.

Examples of Methods

1. **Fairness Constraints:** Use legal standard for anti-discrimination in machine learning algorithms.
2. **Evaluation Metrics:** Include evaluation metrics other than accuracy.

IN-PROCESSING METHODS

FAIRNESS CONSTRAINTS

DESCRIPTION

Intuition

- We can create constraints that capture whether specific fairness metrics are being respected by the algorithm.
- Then, if we ensure that the objective function is optimised subject to the given constraint, we can eliminate discrimination as defined by this fairness metric.

How it works

- Labor law example: labor law uses the notion of disparate treatment and disparate impact to try to identify instances of employer discrimination.
    - To eliminate disparate treatment: want our model to optimise the objective function subject to a constraint that ensures that given ie. gender, the probability of a certain output label given the model has access to all features, including gender is equal to the probability of a certain output label given the model only has access to features other than gender.
- How these constraints are applied is algorithm-specific.

EXAMPLE

Use Case

- The method of optimization with a constraint is used in applications such as robotics/control, natural language, and graphical interfaces research.
- Some of the advantages of using this method are: easy problem modeling and constraints provide a natural way of implementing propagation rules.
- Some of the disadvantages are that optimization may not be very effective and there is a risk of over-constraining the model.

To learn more

- Zafar et. al., Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment.
- Leader, Jeffery J. (2004)

IN-PROCESSING METHODS

	DESCRIPTION	EXAMPLE
EVALUATION METRICS	Intuition • If you only look at accuracy as your evaluation metric, you won't be able to understand the true intricacies of your algorithm's predictions. How it works • Look at evaluation metrics other than accuracy. ◦ Examples include precision, recall, F1, and a confusion matrix. ◦ Then, you can look at each error metric when stratified by protected attribute. ◦ This will allow you to determine how being part of a protected class affects the error rate of the algorithm's prediction for that data point.	Use Case • Suppose we have trained an algorithm which assesses the probability that a defendant will become a recidivist and that this algorithm will be used by judges and parole officers to determine sentencing. In this case, it would be particularly concerning to have a very high rate of false negatives, since that would mean non-recidivists would be receiving unfairly long sentences. Given that there is proof of racial bias in these algorithms, it would be helpful to stratify the false negative rates by race in order to see if protected groups have a greater false negative rate than unprotected groups. From there, the algorithm could be tweaked using other methods to correct for this.

RESOURCES FOR ALGORITHM-SPECIFIC IN-PROCESSING METHODS

ALGORITHM	RESOURCES
LINEAR REGRESSION	• Berk et al., A Convex Framework for Fair Regression • Agarwal et. al., Fair Regression: Quantitative Definitions and Reduction-based algorithms
LOGISTIC REGRESSION	• Berk et al., A Convex Framework for Fair Regression • Agarwal et. al., Fair Regression: Quantitative Definitions and Reduction-based algorithms • Zafar et al., Fairness Constraints: Mechanisms for Fair Classification
DECISION TREES	• Grari et al., Achieving Fairness with Decision Trees: An Adversarial Approach • Raff et al., Fair Forests: Regularized Tree Induction to Minimize Model Bias • Aghaei et al,. Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making
RANDOM FOREST	• Raff et al., Fair Forests: Regularized Tree Induction to Minimize Model Bias
SVM	• Zafar et al., Fairness Constraints: Mechanisms for Fair Classification

ACCOUNTING FOR BIAS

MANY APPROACHES FOR ACCOUNTING FOR BIAS FALL INTO ONE OF THREE CATEGORIES

A. PRE-PROCESSING METHODS

B. IN-PROCESSING METHODS

C. **POST-PROCESSING METHODS**

POST-PROCESSING METHODS

FOCUSED ON THE MACHINE LEARNING ALGORITHM AND ITS OUTPUTS

This section focuses on post-processing methods to debias your machine learning algorithm, which are methods that alter the model's internals and predictions after the model has been trained on the dataset.

Examples of Methods

1. **Threshold Tuning:** Tune thresholds to promote or demote predictions that fall close to the decision boundary lines.
2. **Calibration:** Alter probabilities of class membership so that the probabilities are closer to the true likelihood.
3. **Transparency and Explainability:** Ensuring that one is transparent about the assumptions used by their model and that their model's outputs can be explained and understood.

POST-PROCESSING METHODS

	DESCRIPTION	EXAMPLE
THRESHOLD TUNING	Intuition • Adjusting the threshold that determines whether a data instance will receive a positive or negative target label can change the number of false positive and false negative errors. This is especially important in instances when having a false positive or a false negative is particularly consequential and an unprotected group has a greater number of false positive or false negative instances than the protected group. How it works • Try different threshold values and look at how changing the threshold affects precision and recall. Pick an appropriate threshold based on these observations.	Use Case • Useful for machine learning algorithms that output a real number and classify the data instance as positive or negative class based on whether the real number falls above or below a threshold. ◦ Example: logistic regression To learn more • Hardt et al., Equality of Opportunity in Supervised Learning
CALIBRATION	Intuition • Algorithms that predict pseudo-probabilities or probabilities are often over-confident or under-confident in their predictions. • If probabilities are systematically off-mark for a protected group, can calibrate the probabilities so that it is closer to the true probability for that protected group. How it works • Attempt to fit distribution of predicted probabilities to distribution of probabilities observed in the training data using a function.	Use Case • Useful for machine learning models that predict pseudo-probabilities or probabilities of class membership. ◦ Example: SVMs, logistic regression, naive bayes, deep neural networks with final softmax layer To learn more • Pleiss et al., On Fairness and Calibration

POST-PROCESSING METHODS

	DESCRIPTION	EXAMPLE
TRANSPARENCY AND EXPLAINABILITY	Intuition • To mitigate the risk of black box AI, transparency in algorithmic operation is being demanded. This includes knowledge of how these algorithms operate in practice and throughout the entire workflow in which models are built, trained, and deployed. Innovative frameworks for algorithmic transparency—also known as explainability, interpretability, or accountability—are gaining adoption among working data scientists. Explainability is important, as it allows people to assess how much developers are doing to mitigate bias. How it works • Black box models are created directly from algorithmic data. Humans cannot understand how variables are being combined to make predictions. Even with a list of the input variables, black box predictive models can be such complicated functions that no human can understand how the variables are related to each other and reach a final prediction. • Interpretable models, which provide a technically equivalent, but possibly more ethical alternative to black box models, are different. These models are constrained to provide a better understanding of how predictions are made.	Use Case • Explainability is crucial in every AI algorithm. It should also be considered in every application with decisions that impact people.

PART 3: MITIGATING ALGORITHMIC BIAS

“There’s a real danger of systematizing the discrimination we have in society [through AI technologies]. What I think we need to do — as we’re moving into this world full of invisible algorithms everywhere — is that we have to be very explicit, or have a disclaimer, about what our error rates are like.

-Timnit Gebru, Research Scientist, Google

ETHIXAI
ALGORITHMIC AUDIT

Try out recommendations in this checklist to help guide your journey to de-biasing your Tech product. not every recommendation will apply to your product, but here are some suggestions to get you started.

PREPROCESSING RECOMMENDATIONS

CONSIDERATIONS TO DECREASE LABELLING BIAS

- □ Alter class labels of selected instances that fall on the decision boundary to increase representation of protected groups with "favorable" label. See page 16.
- □ Assign different weights to instances based on their group membership so that for example protected groups with "favorable" label have greater weight. See page 16.
- □ Ensure multiple people label each instance and that people labelling have diverse demographics to limit implicit biases. Is your labeling criteria explainable?
- □ Think critically about the source of your dataset. Are certain groups overrepresented in your dataset?

CONSIDERATIONS TO DECREASE SAMPLING BIAS

- □ Check that no group is over or underrepresented based on the population you are attempting to represent.
- □ Increase representation of protected groups in dataset by oversampling and undersampling certain data instances

IN-PROCESSING RECOMMENDATIONS

In-Processing techniques are algorithm-specific, so research in-processing techniques to mitigate bias specific to your chosen learning algorithm.

□ Research in-processing techniques to mitigate bias specific to the learning algorithm you are using.

□ Look into whether using different fairness constraints is a useful technique for the problem you are trying to optimize. See table on page 23 for some constraint ideas for different algorithms.

□ Use multiple evaluation metrics that reveal whether predictions vary when looking only at protected vs unprotected groups. See page 22.

POST-PROCESSING RECOMMENDATIONS

Post-processing techniques must be an iterative process. As novel data is consumed by your learning algorithm, it may develop new biases, which makes constant monitoring imperative.

IMPLEMENTATION OF PROCESSES FOR ONGOING MONITORING

□ Identify and implement a process for ongoing monitoring of your algorithm for potential bias.

□ Consider hiring a third-party to check for and monitor potential biases in your learning algorithms over time.

CONSIDERATIONS TO DECREASE BIAS OF PREDICTIONS

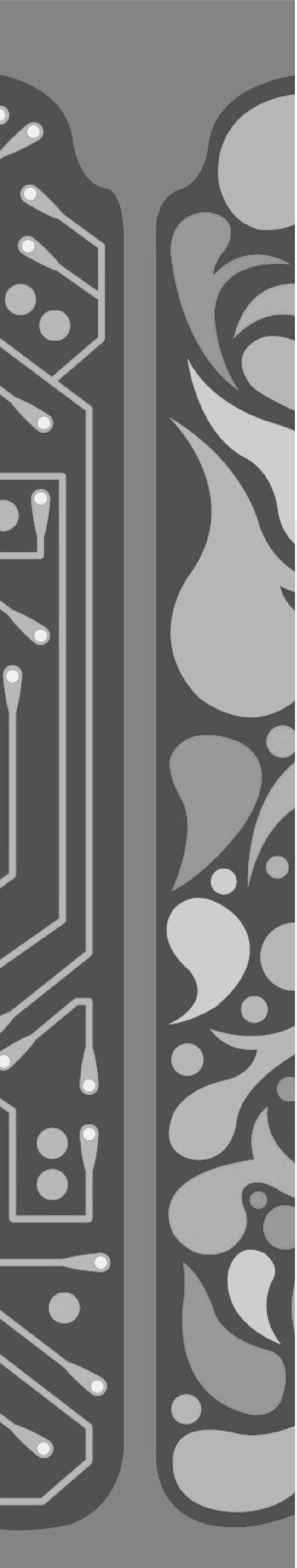
□ Tune thresholds to promote or demote predictions that fall close to the decision boundary lines.

□ Alter probabilities of class membership so that the probabilities are closer to the true likelihood.

CONSIDERATIONS RELATED TO EXPLAINABILITY & TRANSPARENCY

□ Be transparent about the assumptions used by your learning model.

□ Be able to explain model's outputs so that a human can understand the decision-making process.



PART 4: CASE STUDIES AND APPLICATIONS

“Unfortunately, we have biases that live in our data, and if we don’t acknowledge that and if we don’t take specific actions to address it then we’re just going to continue to perpetuate them or even make them worse.”

- Kathy Baxter, Ethical AI Practice Architect, Salesforce

BIAS IN CRIMINAL JUSTICE

Parole officers, judges and probation officers in the US are using machine learning algorithms to assess how likely a defendant is to commit another crime. There are many risk assessment machine learning algorithms in use, some of which are built for a particular state, while others are created by people in academia. COMPAS, which stands for Correctional Offender Management Profiling for Alternative Sanctions, is a commercial tool created by Northpointe, Incorporated.

ProPublica investigated COMPAS in 2016. ProPublica's analysis found that white defendants were favored considerably more than black defendants. According to the COMPAS system, black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism. Simultaneously, white defendants were more likely than black defendants to be incorrectly flagged as low risk for committing another crime.

Further, police across the US have used facial recognition to supplement their work on the ground. Recently, studies by M.I.T. and the National Institute of Standards and Technology have raised concerns about these systems. While facial recognition technology works well on white men, the results are far less accurate for other demographics. Given these inconsistencies, Amazon, Microsoft and IBM have recently decided that they would stop or pause their facial recognition offerings for law enforcement. However, it is important to note that companies such as Vigilant Solutions, Cognitec, NEC, Rank One Computing and Clearview AI still supply the vast majority of law enforcement agencies with their surveillance technology.

BIAS IN HEALTH CARE

A machine learning algorithm widely used by health management systems to select patients for "high-risk care management" programs predicts risk scores for patients each year. Those with the highest risk score get selected for these programs and are provided with more resources and medical attention than those in, for example, the 55th percentile who will simply be referred to a primary care physician. It was found that this algorithm has racial biases, since instances of less-healthy Black patients who should receive more medical attention scored similarly to instances of more-healthy White patients, as a result of which they received less medical attention. This has life-threatening consequences since it leads to white patients receiving more care than black patients.

BIAS IN HIRING

Hiring algorithms are susceptible to racial biases. Human resource managers face the task of evaluating large pools of applicants so resume-scanning algorithms are often used to weed out nearly 80% of applicants. In theory, this allows human resource employees to focus on the top 20% of applicants. However, as resume scanners are often trained on past company successes, they are subject to pre-existing inherited biases.

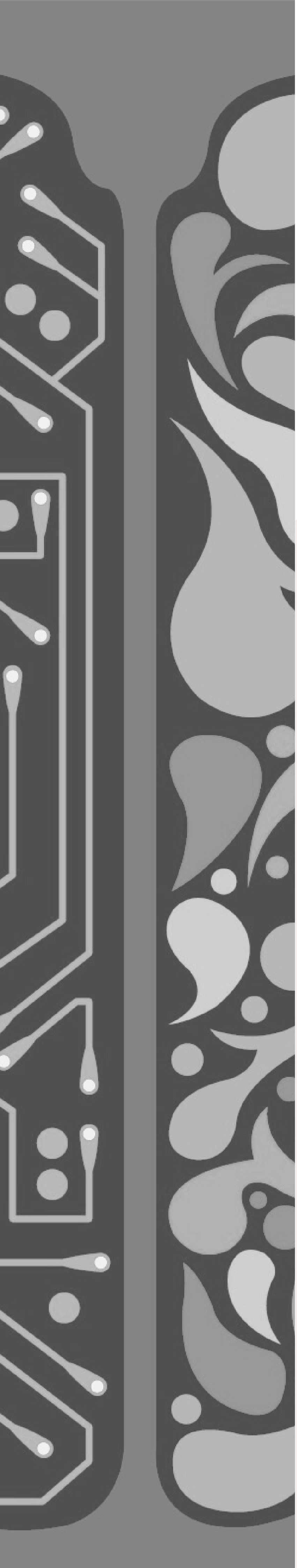
e.g.
If a company does not have a history of hiring people of color, the correlation between people of color and successful employees will be weak as people of color are not fairly represented in the dataset. In this way, the training data that they use to create the hiring algorithms will be biased**.** In this case, machine learning algorithms create black boxes containing racist biases yet are marketed as objective tools to help busy human resources employees.

Amazon Inc.'s algorithm for vetting resume's was trained on a dataset of past applicants, which consistent mostly of male applicants. For this reason, their model was biased against female applicants when the system went live.

BIAS IN FINANCE

Under the Equal Credit Opportunity Act, Credit scoring in the United States has a long-established right to explanation. Creditors are required to provide applicants with specific reasons for financial decisions that impact them. For example, if someone's loan application is denied, the lender needs to provide an explanation. This regulation also applies to scores like the credit score, where the components of criteria for the score needs to be transparent.

In November 2019, the algorithm of Apple's new credit card offered women lower lines of credit than men even with the same financial information (e.g. even among heterosexual married couples with the same finances, the woman was offered a lower line of credit than the man). The algorithm did not explicitly ask for gender, however just removing an input does not make an algorithm blind to that factor because there are many ways to predict gender.



PART 5: FINAL THOUGHTS

"What all of us have to do is make sure we are using AI in a way that is for the benefit of humanity, not to the detriment of humanity"

-Tim Cook, CEO of Apple

...artificial intelligence algorithms are both disastrous tools and marvels of human creation

Although there are numerous methods for mitigating bias, members of the AI community have not agreed on a robust approach for mitigating bias in data-driven machine learning systems. There are no conclusive results regarding which method in the pre-processing, in-processing and post-processing categories is most effective for a broad set of applications. A thorough and holistic evaluation of the existing methods is necessary to understand their capabilities as well as their shortcomings.

Further, mitigating bias in your AI models *must* be an ongoing process. Despite the lack of a common set of methods that you can use to continuously de-bias your algorithm, following are a few concrete things you should do:

1. Add checks to periodically monitor whether your algorithm contains any unfair biases (spoiler: it always will!) and react accordingly to new information.
2. Recruit a third party to regularly check that your product is not discriminatory towards a protected group.
3. Make de-biasing your machine learning models a company priority.

Like many modern things, AI algorithms are both disastrous tools and marvels of human creation. What matters most is how we create these computational agents, who we include in the development process, and our collective will to shift our intrinsic biases and cultural perspectives. Everyone must continue educating themselves on this topic and advocate for the responsible use of artificial intelligence in both passion projects and in the workplace.

SOURCES

PART 1: A BRIEF INTRODUCTION TO BIAS IN AI

Handbook of Research on Teaching with Virtual Environments and AI, by Gianni Panconesi and Maria Guida, Information Science Reference, 2021, pp. 615–615.

Mitchell, Tom M. Rutgers University, 1980, The Need for Biases in Learning Generalizations.

Geman, Stuart, et al. “Neural Networks and the Bias/Variance Dilemma .” Neural Computation, vol. 4, no. 1, 1 Jan. 1992, pp. 1–58., doi: https://doi.org/10.1162/neco.1992.4.1.1.

PART 3: ACCOUNTING FOR BIAS

3A. PRE-PROCESSING METHODS

Kusner, M., & Loftus, J., & Russell, C., & Silva, R. (2017) Counterfactual Fairness, 31st Conference on Neural Information Processing Systems, Long Beach, CA, 2017

Calders, T., & Kamiran, F., & Pechenizkiy, M. (2009). Building Classifiers with Independency Constraints, IEEE International Conference on Data Mining Workshops, Miami, FL, 2009, Netherlands: Eindhoven University of Technology

Silva, Ricardo, et al. “Counterfactual Fairness.” The Alan Turing Institute, - [ ] www.turing.ac.uk/research/research-projects/counterfactual-fairness.

Kamiran, F., & Calders, T. (2011, November 16) Data Preprocessing Techniques for Classification Without Discrimination. Springer Science and Business Media, vol. 33, no. 1, doi:10.1007/s10115-011-0463-8

SOURCES

3B. IN-PROCESSING METHODS

Zafar, M., & Valera, I., & Rodriguez, M., & Gummadi, K. (2017). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment , International World Wide Web Conference Committee, Perth, Australia, April 3-7, 2017.

Berk, R., & Heidari, H., & Jabbari, S., & Joseph, M., & Kearns, M., & Morgenstern, J., & Neel, S., Roth, A. (2017, June 9). A Convex Framework for Fair Regression. University of Pennsylvania.

Agarwal, A., & Dudík, M., & Wu, Z. (2019). Fair Regression: Quantitative Definitions and Reduction-based Algorithms, 36th International Conference on Machine Learning, Long Beach, California, 2019. PMLR.

Zafar, M., & Valera, I., & Rodriguez, M., & Gummadi, K (2017). Fairness Constraints: Mechanisms for Fair Classification, Fort Lauderdale, Florida, USA, 2017. JMLR.

ZGrari, V., Ruf, B., Lamprier, S. et al. Achieving Fairness with Decision Trees: An Adversarial Approach. Data Sci. Eng. 5, 99–110 (2020). https://doi.org/10.1007/s41019-020-00124-2

Raff, E., & Sylvester, J., & Mills, S. (2018). Fair Forests: Regularized Tree Induction to Minimize Model Bias. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, doi: 10.1145/3278721.3278742.

Aghaei, S., & Azizi, M., & Vayanos, P. (2019). Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making. Association for the Advancement of Artificial Intelligence, doi: 10.1609/AAAI.V33I01.33011418.

3C. POST-PROCESSING METHODS

ZHardt, M., & Price, E., & Srebro, N (2016). Equality of Opportunity in Supervised Learning, 30th Conference on Neural Information Processing Systems, Barcelona, Spain, 2016.

Pleiss, G., & Raghavan, M., & Wu, F., & Kleinberg, J., & Weinberger, K. (2017). On fairness and calibration, Proceesings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, 2017.

SOURCES

PART 4: CASE STUDIES AND APPLICATIONS

Dastin, Jeffrey. “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women.” Reuters, 10 Oct. 2018.

Knight, Will. “The Apple Card Didn't 'See' Gender—and That's the Problem.” Wired, 19 Nov. 2019, www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/ .

Obermeyer, Z., & Powers, B., & Vogeli, C., & Mullainathan, S. “Dissecting racial bias in an algorithm used to manage the health of populations.” Science, vol. 366, issue 6464, DOI 10.1126/science.aax2342

Angwin, Julia, et al. “Machine Bias.” ProPublica, 23 May 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.