An Interactive Framework to Evaluate Algorithmic Discrimination

Aideen Farrell

Universitat Pompeu Fabra

Masters in Intelligent Interactive Systems

Advisor: Carlos Castillo

An Interactive Framework to Evaluate Algorithmic Discrimination

September 2020

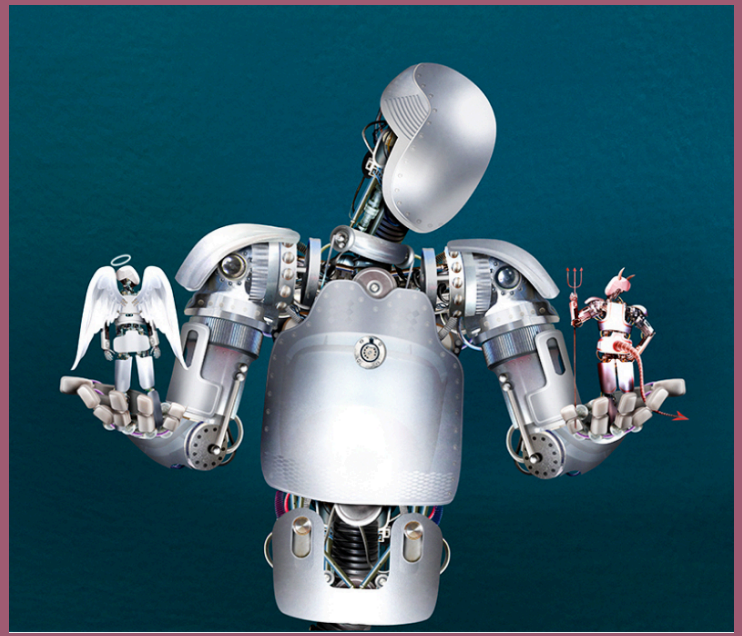
Aideen Farrell

Universitat Pompeu Fabra

Intelligent Interactive Systems

Carlos Castillo

An Interactive Framework to Evaluate Algorithmic Discrimination



Introduction

Introduction – Algorithmic Decisions

Background on algorithmic decisions

Algorithmic Decisions

- Machine learning (ML) systems become more and more part of our daily lives, making predictions and decisions about our potential abilities and trustworthiness.
- The decision could be a simple yes/no, or a score intended to rank our potential (to succeed, in a job, an educational setting, as a law-abiding citizen, in the repayment of a loan, a mortgage, insurance, healthcare and so on).
- Successful learning is based on the assumption that patterns will be repeated; therefor, there is an intrinsic reliance on the past.

Introduction - Algorithmic Discrimination

Background on algorithmic discrimination

Algorithmic Discrimination

- High level definition of bias: *"Any case where AI/ML systems perform differently for different groups in ways that may be considered undesirable." [1]*
- Data gathered from the world has centuries of historical context. The world has centuries of historical prejudice.
- Models that 'learn' from this data often cannot avoid picking up on the patterns of discrimination.
- Representation of specific groups within the data is also influenced by this historical prejudice.

Introduction – Problem

Importance of the problem

Problem

The tools necessary to develop, train and deploy ML models are becoming increasingly accessible.

The possibility of encountering an ML model created with no oversight from a domain expert or data scientists becomes increasingly likely.

Even with oversite there are many fairness definitions, many ways of measuring fairness and many mitigating strategies to choose from.

Deciding upon an organisations fairness strategy requires collaboration and an intentional awareness of consequences in order to align strategy with company values.

These conversations are often quite complex due to conflicting definitions and a heavy focus on math. At the same time there can be serious consequences for individuals, groups and society as a whole.

Objectives

Thesis Objectives

1. Facilitate the detection of bias at various points in Machine learning pipeline.

2. Encourage inter-organisation reflection and collaboration as regards to the dataset, the purpose of model, the real-world implications.

3. Use quantitative analytical and statistical techniques for bias detection and measurements of fairness/unfairness

4. Facilitate a level of abstraction between mathematical measurements and natural language interpretations so that the focus of any conversation is on fairness, not math, to help bridge the gap between technical and non-technical stakeholders and policy makers.

Related Work

Related Work

Focus of Literature Reviews

01 Survey on Industry Practitioner needs.	Improving fairness in machine learning systems: What do industry practitioners need?(2019) [1]
02 Survey on Bias and Fairness.	A Survey on Bias and Fairness in Machine Learning. (2019)[2]
03 Analysis of Fairness Definitions	Fairness Definitions explained(2018)[3]
04 Group v's Individual Fairness	On the apparent conflict between individual and group fairness.(2020)[4]
05 Worldviews	On the (im) possibility of fairness(2016)[5]
06 'Mathiness'	Troubling trends in machine-learning scholarship.(2019)[6]

Related Work

Challenges facing commercial teams, product teams, development teams [2]

Improving fairness in machine learning systems: What do industry practitioners need?(2019)

Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M. & Wallach, H.

A systematic investigation of challenges facing stakeholders in commercial ML development.

- 35 semi-structured interviews across 25 product teams from 10 companies.
- Anonymous survey of 267 ML practitioners.
  - Development of statistical definitions of fairness/unfairness.
  - Algorithmic methods to assess and mitigate each definitions.

- **Practitioners challenges:**
  - Practitioners see data collection as most important while research focuses on mitigation at model, not the data.
  - Research has focused on statistical definitions of fairness/unfairness
  - Framework designs are currently driven by algorithmic methods not by real-world needs.
  - Inter organisation bias 'blind spots' (user study participants, product team etc) have created a need for a more holistic collaborative approach.
  - There is a need for facilitation of alignment and the necessity for collaborative discussions

Related Work

A Survey on Bias and Fairness in Machine Learning [1]

A Survey on Bias and Fairness in Machine Learning. (2019)

Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A.

A survey of real-world applications that have shown biases.

- Identification of twenty three different definitions of the sources of bias that may impact ML applications.
- Identification of six definitions of the different types of discrimination.
- A taxonomy of ten definitions that ML researchers use to describe fairness
- Concluding that lack of synthesis in the definitions creates an impossibility of understanding.
- Conclude that detection of bias in the dataset is an open research problem.

Related Work

Fairness Definitions explained **[3]**

Fairness Definitions explained.(2018)

Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M. & Wallach, H.

An Analysis of Fairness Definitions

- Over twenty definitions of fairness identified.
- No clear definition to understand what to apply within a particular.
- Differences are difficult to grasp.
- Conflict in definitions such that a scenario can be considered fair according to one mathematical definitions and unfair according to another.
- Statistical methods are insufficient alone, due to conflicting notions of fairness.
- Non-Statical notions require domain expert.

Definition
Group fairness or statistical parity
Conditional statistical parity
Predictive parity
False positive error rate balance
False negative error rate balance
Equalised odds
Conditional use accuracy equality
Overall accuracy equality
Treatment equality
Test-fairness or calibration
Well calibration
Balance for positive class
Balance for negative class
Causal discrimination
Fairness through unawareness
Fairness through awareness
Counterfactual fairness
No unresolved discrimination
No proxy discrimination
Fair inference

Related Work

On the apparent conflict between individual and group fairness. **[4]**

On the apparent conflict between individual and group fairness. (2020)

Binns. R

"The apparent conflict between individual and group fairness is more of an artifact of the blunt application of fairness measures, rather than a matter of conflicting principles. In practice, this conflict may be resolved by a nuanced consideration of the sources of `unfairness' in a particular deployment context, and the carefully justified application of measures to mitigate it"

Related Work

On the (im)possibility of fairness[5]

On the (im) possibility of fairness (2018)

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian

- Quantifiable features of a person do not necessarily reflect the *true* features of a person
- Mathematical expression for two world views:

  ***What you see is what you get (WYSIWYG)*:** Whatever the data reflects constitutes an objective picture of the world, even if inequalities appear in the results. The second ‘

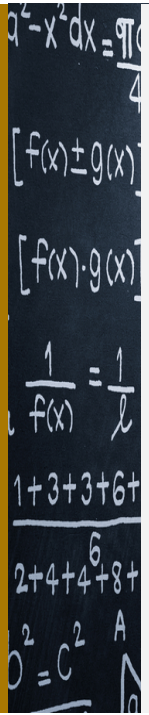
  We’re all equal (WAE): Data is not inherently objective. data does not account for the subjective, bias entrenched influence under which the data was created. WAE, therefore, entails that there be a mathematical compensation for any significant discrepancies in output.

Related work

[Troubling trends in machine learning scholarship[6]

Troubling trends in machine-learning scholarship (2019)

Lipton, Z. C., & Steinhardt, J.



Identification of four observed patterns in the academics of machine learning.

Two of which compound the challenges faced by those concerned with fair ML and with collaboration between data scientists/engineer and the broader collection of stakeholders.

- *Mathiness:* heavy math based discussions around fairness often obfuscates the *real-world* point of view.
- *Misuse of language:* complex legal doctrine is often appropriated leading to confusion.

Methodology

The Problem to solve

Detect Disparities

Data is often unrepresentative of the ground truth, not reflecting true demographic representation and/or reflecting existing prejudices. This can often have serious real world consequences for individuals, groups and society as a whole.

Collaborate/Align

While decisions are being made for various business reasons other than fairness or discrimination, multiple fairness definitions exist for which maths can be used to justify any goal.

It is important that all stakeholders are aligned in understanding and values.

Less “Mathiness”

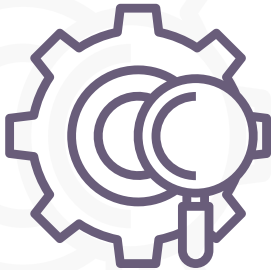
To facilitate alignment mathematical/Statistic based measures of fairness should have a natural language interpretations as heavy math based discussions around fairness often obfuscates the *real-world* point of view.

Methodology

Qualitative and Quantitative.

Qualitative research was used used both to define the problem and to develop and approach to the problem. This research was focused on understanding the discussions highlighted in the related work, particularly the surveys on practitioner needs which interviewed 35 practitioners, across teams and 10 companies in addition to an anonymous survey of 267 industry ML practitioners.

Use Case involves both qualitative and quantitively aspects. The framework has been modelled based on qualitative research however the use case results are of a quantitative nature, based on statistical, mathematical and theoretical propositions and measurements.



Process

Steps in research process

Analysis of Related work

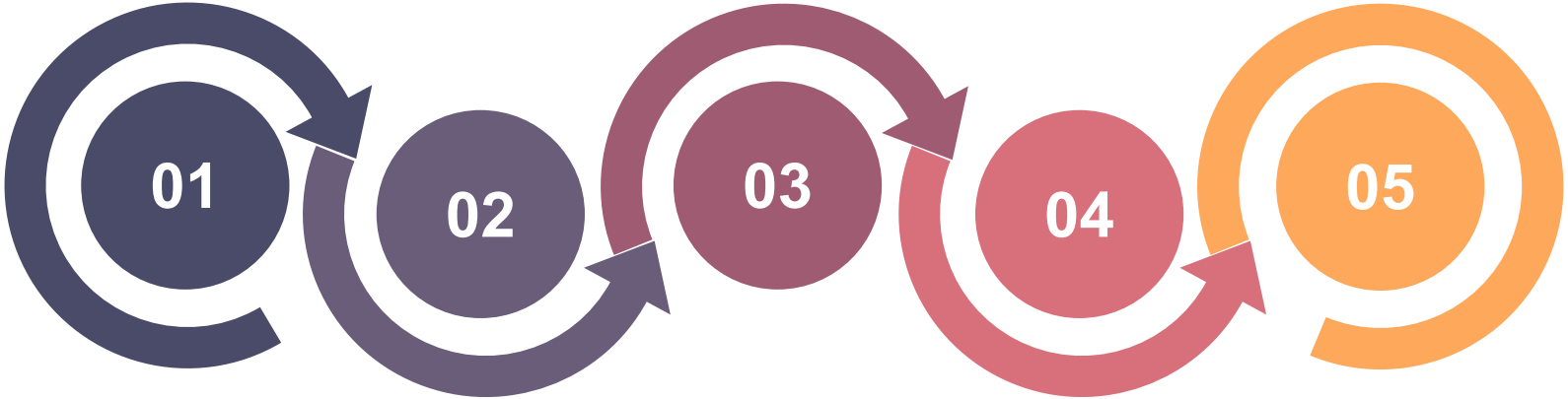
Practitioner needs.
Definitions of Fairness.
Concept of worldview.
Bias measurement strategies.
Counterfactual Fairness/ Counterfactual Thinking

3. Model Fairness Approach

Investigated several API's and decided upon **Aequitas** due to the extensive functionality and documentation.

5. Ranking Fairness Approach

Standard logarithmic discount using average.
Standard logarithmic discount using sum.
Geometric Distribution.
FA*IR.



2. Data Analysis Approach

Identified the types of bias that framework will detect.
Reflective questions to ask
Statistic test
Distributions across groups
Correlations.

4. Model Explainability Approach

Shap: Classic Shapley values from game theory.
Model Agnostic option.
Well documented

Process

Identify the bias the Framework will address and Methods for addressing it

	Representation Bias	Historical Bias	Measurement Bias	Evaluation Bias	Presentation Bias	Ranking Bias
Data Analysis	How has the population been sampled, Are all groups proportional to their representation in the general population? Proportional to representation In the domain?	What is the distribution of target and features across groups. Statistical significance test for disparities in distribution Correlation between features and protected features.	Measurement bias relates to how we decide to measure a particular feature or target. Perhaps through use of a proxy. How do we define 'risky' 'at risk' 'intelligent', 'homeless.	What are the real world consequences of the outcome, Is it assistive or punitive? Are we considering the real world consequences of false positives and/or false negatives?		
Model Fairness			What importance is placed on the various features the model interprets when making a decision	How is the models accuracy being evaluated? Are we considering the punitive vs assistive nature of the outcomes. Are we using group fairness metrics? Has there been any reflection on counterfactual		

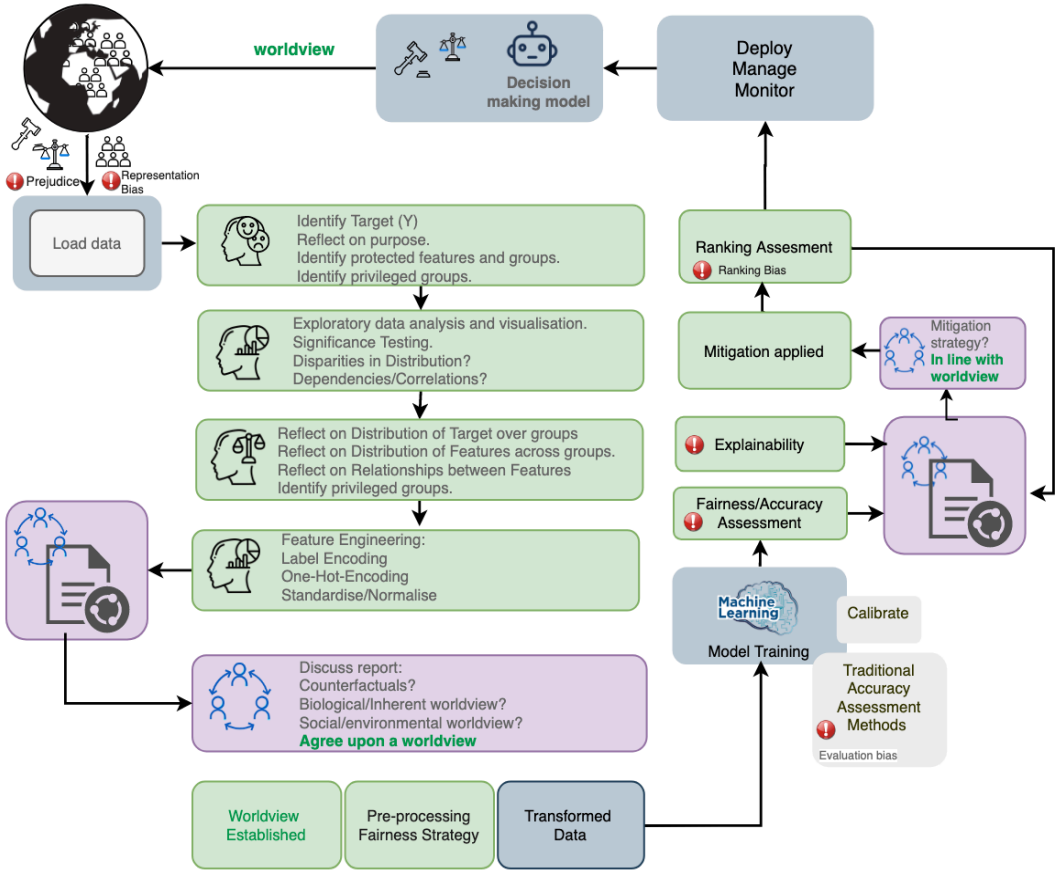
Process

Identify the bias the Framework will address and Methods for addressing it

	Representation Bias	Historical Bias	Measurement Bias	Evaluation Bias	Presentation Bias	Ranking Bias
Model Explainability	How has the population been sampled, Are all groups proportional to their representation in the general population? Proportional to representation In the domain?		Measurement bias relates to how we decide to measure a particular feature or target. Perhaps through use of a proxy. How do we define 'risky' 'at risk' 'intelligent', 'homeless.	What are the real world consequences of the outcome, Is it assistive or punitive? Are we considering the real world consequences of false positives and/or false negatives? Are measurements compared across groups.		
Ranking				How is the ranking being evaluated. Are measurements compared across groups.	What is the distribution of the score /ranking across groups. Statistical significance test. What is the purpose of application, will all results be considered?	**Exposure per group:** Standard logarithmic discount using average Standard logarithmic discount using sum. Geometric Distribution FA*IR

Framework definition

Proposed workflow for bias reflection framework.



Use Case

Use case – Law School dataset

The Law school data

The law school data set: The Law School Admission Council surveyed 163 law schools in the U.S. **21,790** rows of data from 1991. Students who entered law school in Autumn 1991, tracked for 3 years and 5 rounds of the students' BAR examinations.

01 **UGPA** (Undergraduate grade point average): The undergraduate grade-point average collected before law school.

LSAT (Law school admissions test): The Law School Admission Test is an integral part of law school admission in the United States, Canada and several other countries.

03 **Sander_index**: Sander proposes combining the LSAT and GPA scores into a single weighted average, using weights that correspond roughly to those used in many law schools. This average is referred to as the "Sander Index.

04 **ZFYA:** Measure of first-year academic achievement in law school, presented as a z-score.

Protected **Race**

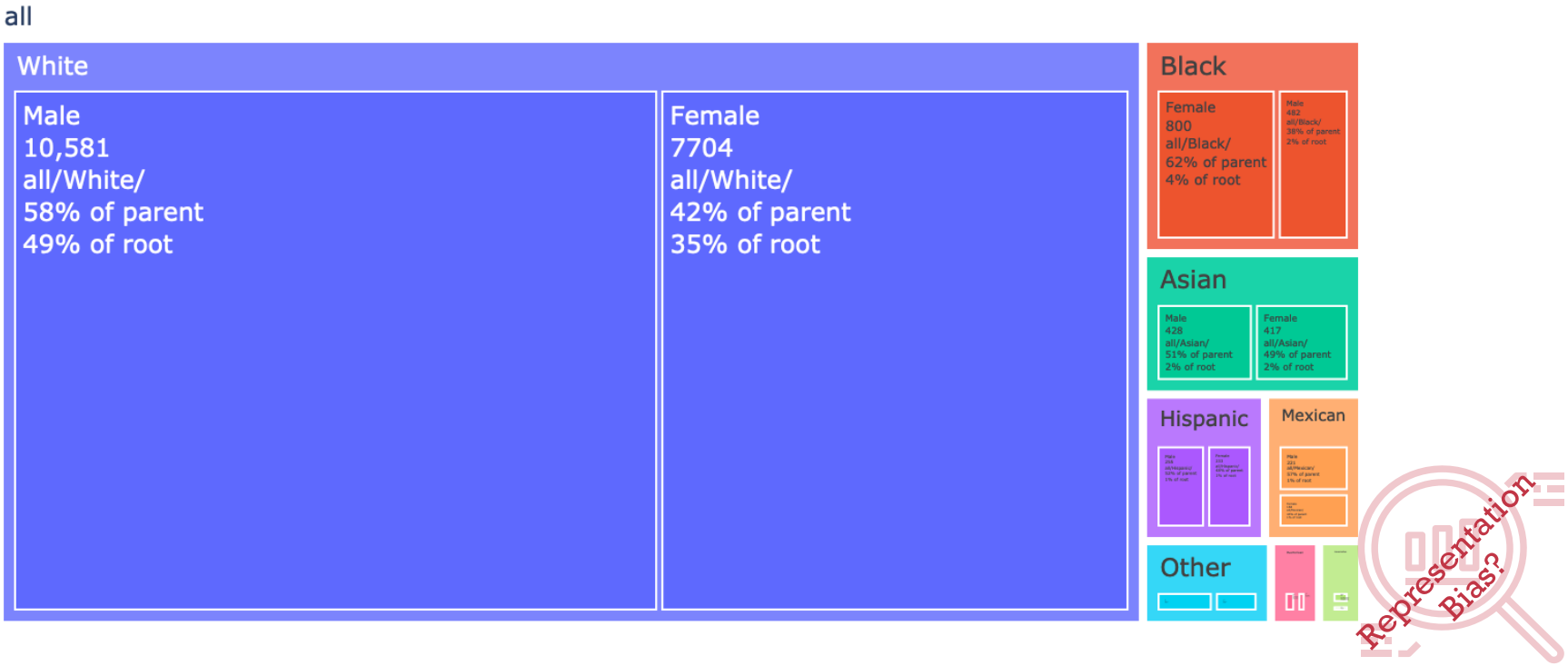
Protected **Gender**

Use Case - Demo

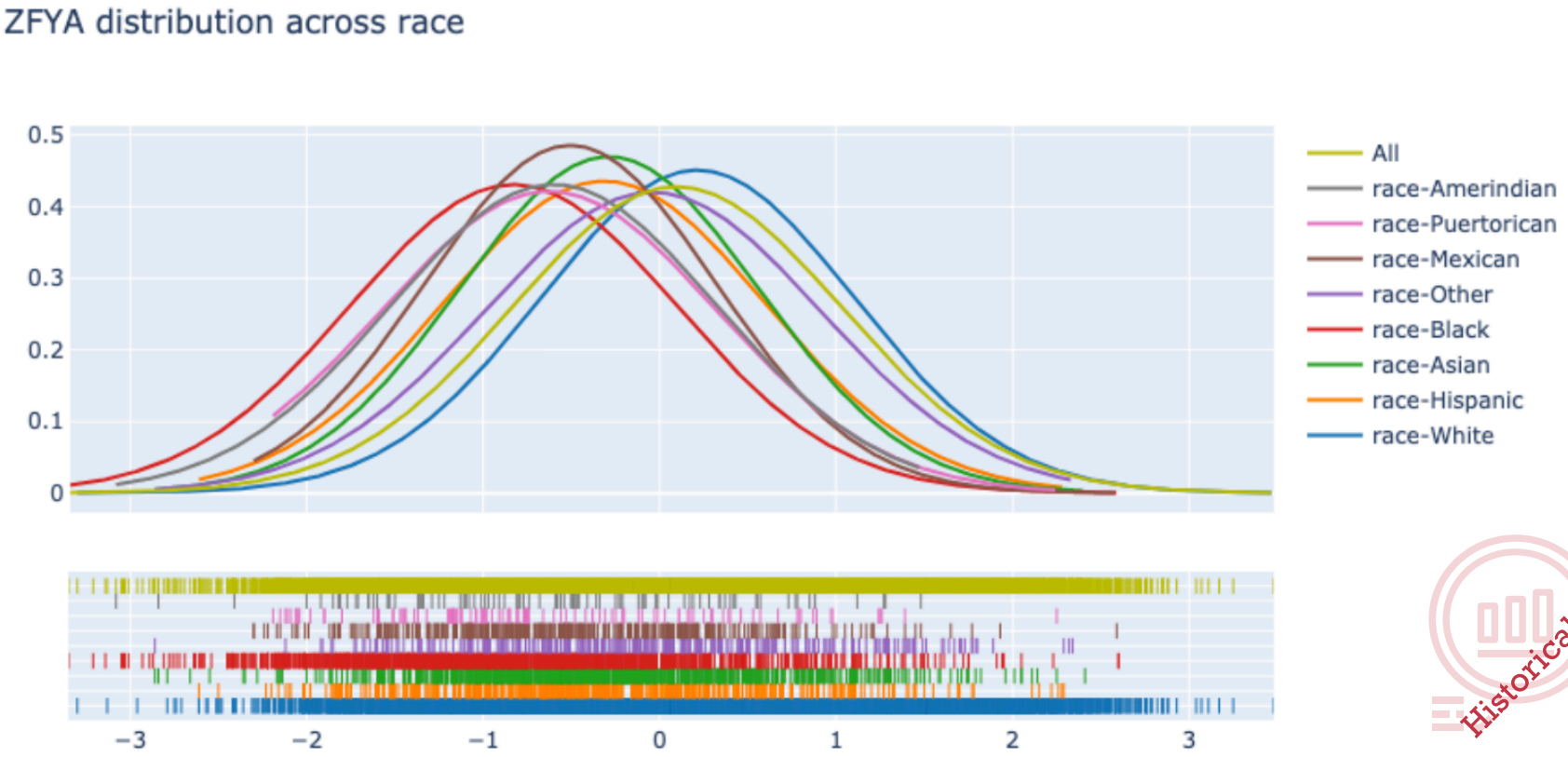
Use Case - Results

Data analysis: Representation of groups

Representation of Protected group(s) in the data



Data analysis: Distribution of outcome



Data analysis: Statistical significance test

Two-Tailed T-Test

Significant variance: The statistic test will tell us if there is a significant variance in the distribution and if this variance is due to chance, or how likely it is that it is not due to chance but rather to an unobserved factor.

T-Value: This value represents the distance between the observed distribution and the expected distribution in a fair world. The larger the value of T, the greater the evidence against the difference occuring by chance in a fair world.

Reference Group: White

Focal Group: Black

	Test Statistic(T-Value)	P-value
Sample Data	40.63077572088865	0.0

There is a 0.0% probability that a difference of 40.63077572088865 occured by chance.

Statistical significance is the likelihood that a relationship between two or more variables is caused by something other than chance.

Data analysis: Review findings

ZFYA: There has been a significant disparity in the distribution of outcome (ZFYA)

A statistically significant difference of T-value **40.6** per the two-tailed T-test.	
White Min: -3.3	Black Min: -3.35
White Max: 3.48	Black Max: 2.6
White Mean: 0.21	**Black Mean: -0.822**
A correlation of **.28** was also observed between race and ZFYA	

UGPA: There has been a significant disparity in the distribution of outcome (UGPA)

A statistically significant difference of T-value **32.2** per the two-tailed T-test.	
White Min: 1.9	Black Min: 1.8
White Max: 4.2	Black Max: 4
White Mean: 3.3	**Black Mean: 2.89**
A correlation of **.20** was also observed between race and UGPA	

LSAT: There has been a significant disparity in the distribution of outcome (LSAT)

A statistically significant difference of T-value **57.7** as per the two-tailed T-test.	
White Min: 17	Black Min: 11
White Max: 48	Black Max: 47
White Mean: 37.5	**Black Mean: 29.4**
A correlation of **.36** was also observed between race and LSAT	

Sander_index: There has been a significant disparity in the distribution of outcome

A statistically significant difference of T-value 61.5 per the two-tailed T-test.	
White Min: .44	Black Min: .38
White Max: 1	Black Max: .91
White Mean: .77	Black Mean: .64
A correlation of .38 was also observed between race and Sander_index	

Review Findings

The analysis of the data has shown significant disparities between Black and White applicants.

Data analysis: Worldview

Collaborate!
Agree on **worldview** in relation to context and to specific feature/target

Reflections related to group representation:

Race

Q: Do you observe any significant disparity in '**race**' group representation between the sample population and the general population in the geographical region of use?

Yes, significant✔	No, not significant

Q: Do you observe any significant disparity in '**race**' group representation between the sample population and the population that the machine learning model will make predictions about after deployment?

Yes, significant✔	No, not significant	Not Applicable

Q: What worldview do you believe should be applied to any significant disparity in '**race**' group representation?. Any disparity in representation is likely caused by:

An inherent characteristic of the protected group✔	An external, unobserved causal influence

Q: In your opinion should this data be used if the objective is to train a fair ML model which will reflect the worldview selected for disparities in **race?**

Yes✔	No	Discussion required

Model Fairness

Proportional parity

Proportional parity: Proportional parity is a representational based group fairness metric which states that each group should have the same proportion of beneficial(non-punative) outcomes. A desire to correct for the absence of proportional parity (when no biological or inherent reason accounts for its' absence) reflects a worldview which recognises the existance of prejudice and a wish to create a "decision maker" willing to apply corrective measures to counter historical discrimination against a particular group or groups and ensure that all groups are proportionately represented in beneficial outcomes. The "decision maker" is aware that such intervention may be reflected in a reduction of perceived utility of *current* model accuracy.
Note These values are calculated based on the group representation in the sample which does not necessarily match that of the population or the domain in which the model will be used.

The privileged group has been set as: White

	Group	Beneficial outcome percentage	Punative outcome percentage
0	Amerindian	0.00	100.00
1	Asian	1.50	98.50
2	Black	0.00	100.00
3	Hispanic	3.74	96.26
4	Mexican	0.94	99.06
5	Other	23.58	76.42
6	Puertorican	0.00	100.00
7	White	68.09	31.91

Proportional Parity.
Demographic Parity.
Equality of Opportunity.
Equality of Outcomes.

Model explainability

SHAP interpretability via Machnamh

(SHapley Additive exPlanations) KernelExplainer is a model-agnostic method which builds a weighted linear regression by using training/test data, training/test predictions, and whatever function that predicts the predicted values. SHAP values represent a feature's responsibility for a change in the model output. It computes the variable importance values based on the Shapley values from game theory, and the coefficients from a local linear regression.

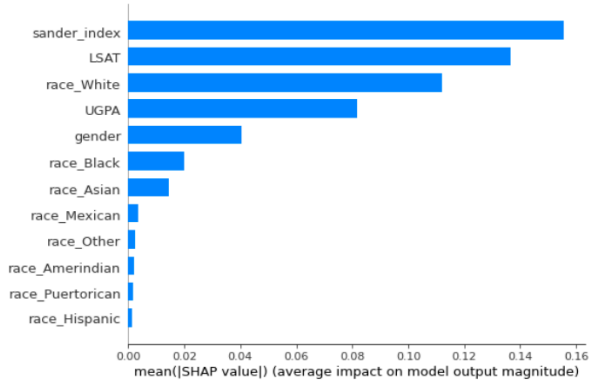
see: https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

It offer a high level of interpretability for a model, through two distinct approaches:

Global interpretability — the SHAP values can show how much each predictor contributes, either positively or negatively, to the target variable. Similar to a variable importance plot however it also indicates the positive or negative relationship between each feature and the target output.

Local interpretability — each observation is assigned it's own SHAP value. This provides a very granular level of transparency and interpretability where we can determine why an individual cases receive a specific prediction and the contribution of each feature to the prediction. Generally speaking variable importance algorithms usually only show the results across the entire dataset but not on each individual case.

Summary Importance | Importance plot | Dependence plot | Individual force plot | Collective force plot



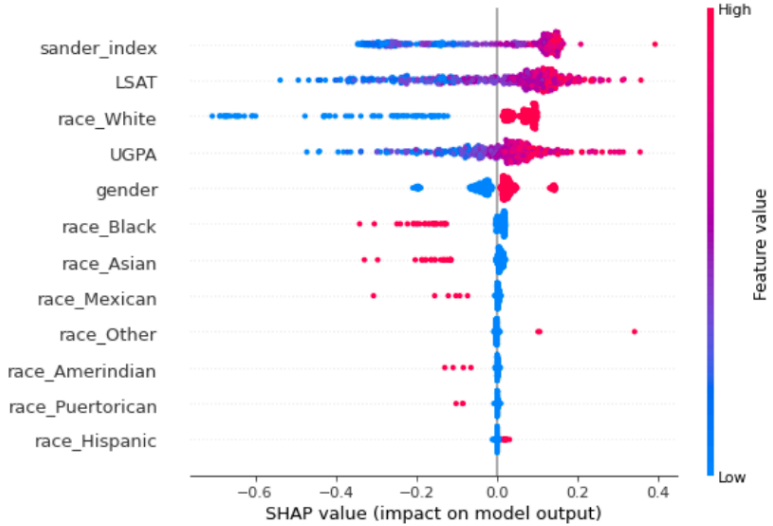
Summary importance plot

Feature importance: Variables are ranked in descending order. The top variables contribute more to the model than the bottom ones and thus have high predictive power.

What input features have the most impact on the decision

Model explainability

Importance plot: lists the most significant variables in descending order. The top variables contribute more to the model than the bottom ones and thus have high predictive power.



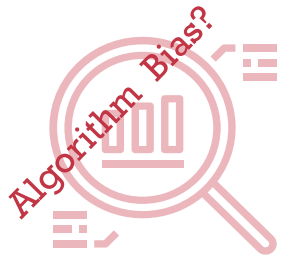
Feature importance: Variables are ranked in descending order.

Impact: Horizontal location indicates if effect of feature is associated with a higher or lower prediction.

Original value: Colour indicates if feature variable is high(red) or low(blue) for the particular observation.

Correlation: A high or low impact(indicated by colour), a positive or negative impact(indicated by position on x-axis)

What is the direction of the impact.



Fairness in ranking

01 Standard logarithmic discount using average

Detect Inequality of attention using a standard logarithmic discount as a measure of exposure drop off, where:
exposure = 1/math.log(j+2, 2)
***j* = position in list.**
Calculate exposure for each position in the ranked list. Compare the **average** of exposure per group against the sum of relevance per group Display 'Group Equity of attention'.

02 Standard logarithmic discount using Sum

Detect Inequality of attention using a standard logarithmic discount as a measure of exposure drop off, where:
exposure = 1/math.log(j+2, 2)
***j* = position in list.**
Calculate exposure for each position in the ranked list. Compare the **sum** of exposure per group against the sum of relevance per group Display 'Group Equity of attention'.

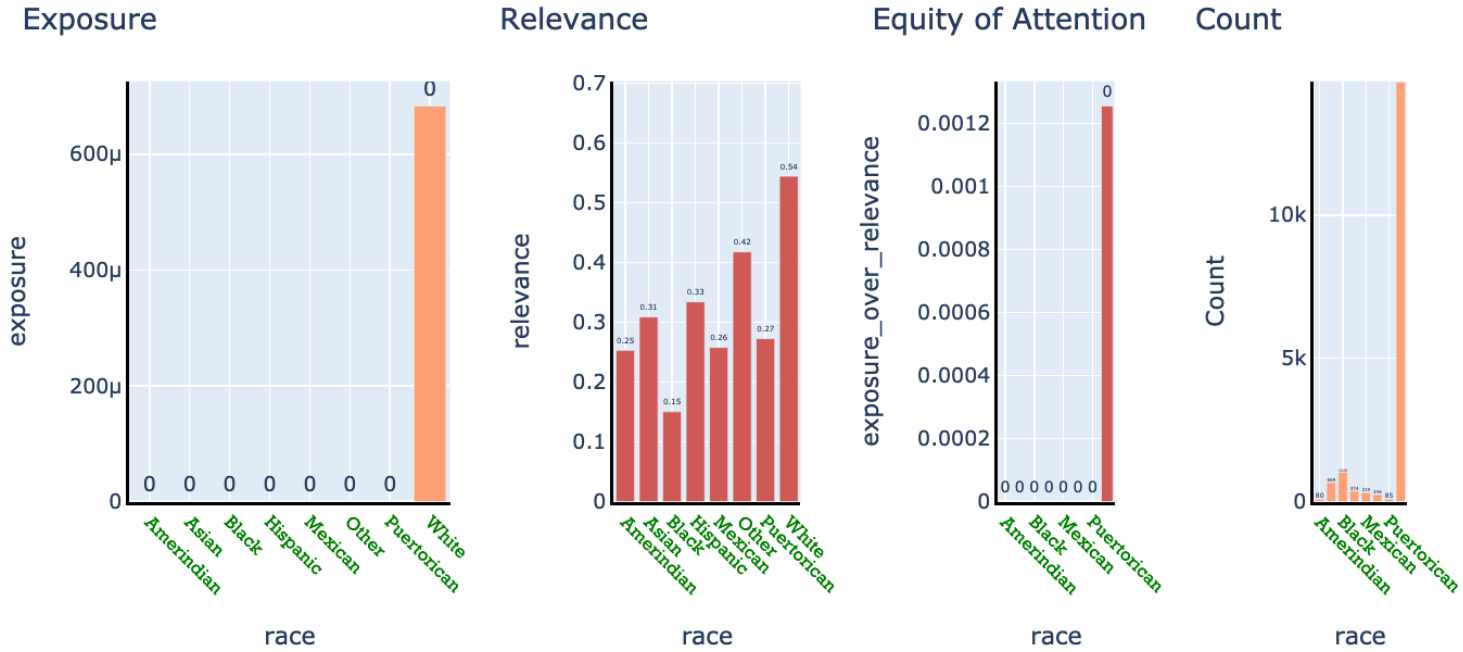
03 Geometric Distribution

Exposure is set to *p* up to a position *k*, and to *0* when the position is lower than *k*. Currently *p* is set to 1 and *k* is set to 10 and group mean value is as displayed.

04 FA*IR

The FA*IR analysis is based on the assumption that the top-k ranked list is fair if it consists of a sufficient proportion of at least p of the protected group for every prefix(n) of the top-k ranking from n = 0 to n = k.

Fairness in ranking



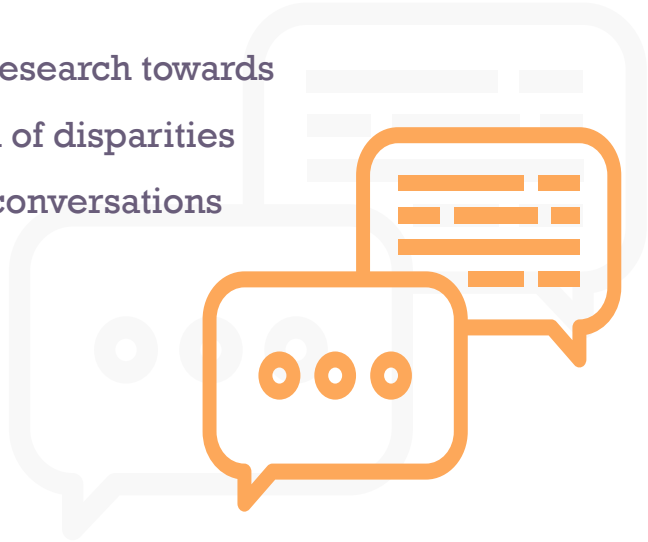
Geometric distribution results

Conclusions

Discussion

This work differs from much of the prior work in that it considers the worldview as the means through which the final mathematical definitions of fairness should be selected, such that the decision is framed through company values rather than math.

It is hoped that this framework can act as a starting point for further research towards the development of a practical framework to facilitates the detection of disparities across groups, invoking reflection and supporting natural language conversations around fairness, supported by but not framed through math.



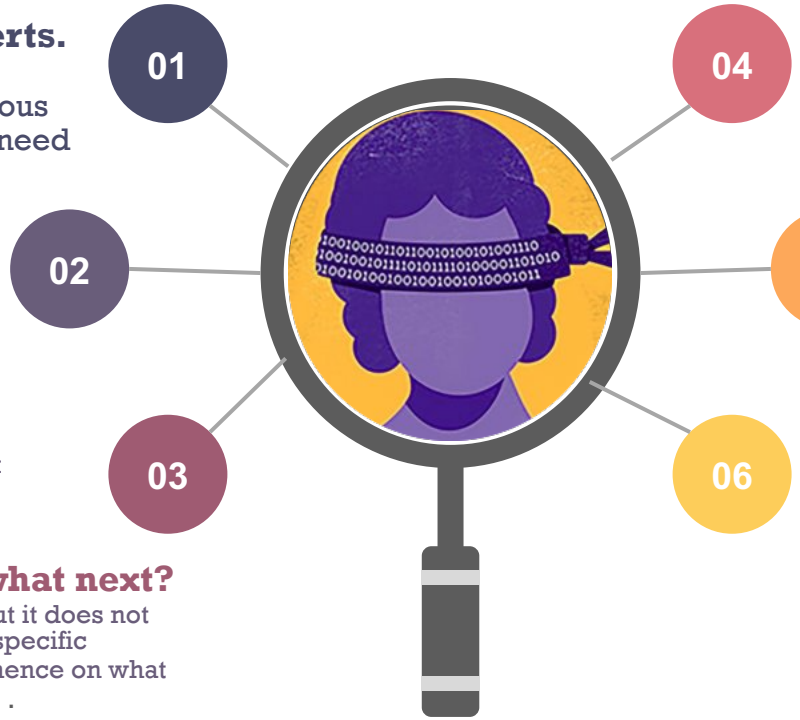
Limitations of Work

Limitations of framework

Lack of input from social science and domain experts.
The worldview definitions and explanations related to the various high level concepts of fairness need to be reviewed.

Not all definitions of bias are addressed!
The framework does not address all of the various definitions of bias. At least 23 types of bias were identified in related work, the framework specifically considers six out of twenty three.

Worldview identified- what next?
The framework provides a focus but it does not solve the overall dilemma of what specific definition of fairness to apply and hence on what mitigation strategy to engage with. .



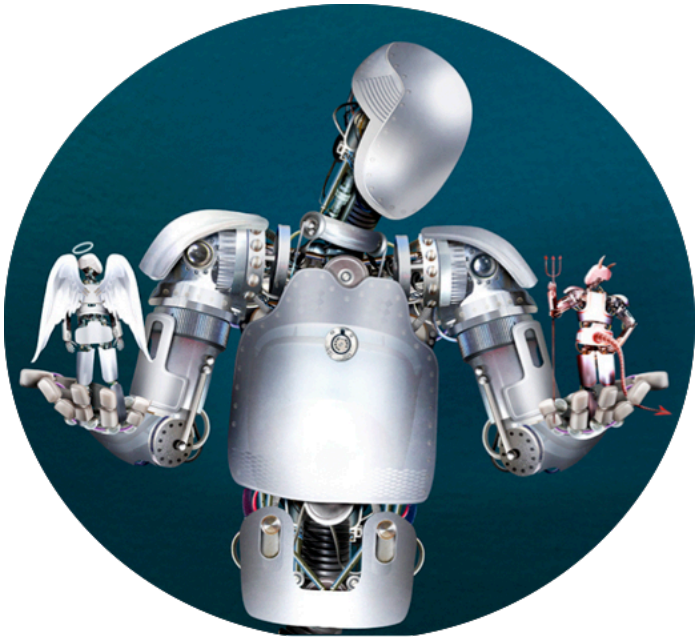
Protected features are not always available!
Sensitive demographics information is not always available.

Limit on algorithms supported!
The framework is restricted to binary classification models, regression models and ranking.

Limited usability testing
Very limited usability testing carried out to determine if features are actually intuitive.

Suggestions

Suggest Future Research



Input from social science domain

There is a need for input from other fields of study in order to re-asses and review the natural language being used to describe the various concepts of fairness and of worldviews.

Mitigation strategy options

The framework is currently solely focused on unfairness detection and the conversation around the objectives of fairness mitigation, however it does not cover specific strategies.

Further research

There is limited research or tools available to facilitate inter-organization alignment , also in the area of analyzing and gathering data particularly when information on protected features may not be available

References

References

1. Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M. & Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? in *Conference on Human Factors in Computing Systems - Proceedings* (Association for Computing Machinery, **2019**). doi:10.1145/3290605.3300830

2. Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A, A Survey on Bias and Fairness in Machine Learning(2019) arXiv:1908.09635

3. Verma, S., & Rubin, J. (**2018**). Fairness definitions explained. In *Proceedings - International Conference on Software Engineering* (pp. 1–7). IEEE Computer Society. https://doi.org/10.1145/3194770.3194776

4. Binns, R. (2020). On the apparent conflict between individual and group fairness. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 514–524). Association for Computing Machinery, Inc. https://doi.org/10.1145/3351095.3372864

5. Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, September 2016.

6. Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine-learning scholarship. *Queue*, *17*(1). https://doi.org/10.1145/3317287.3328534

QUESTIONS

THANK YOU !