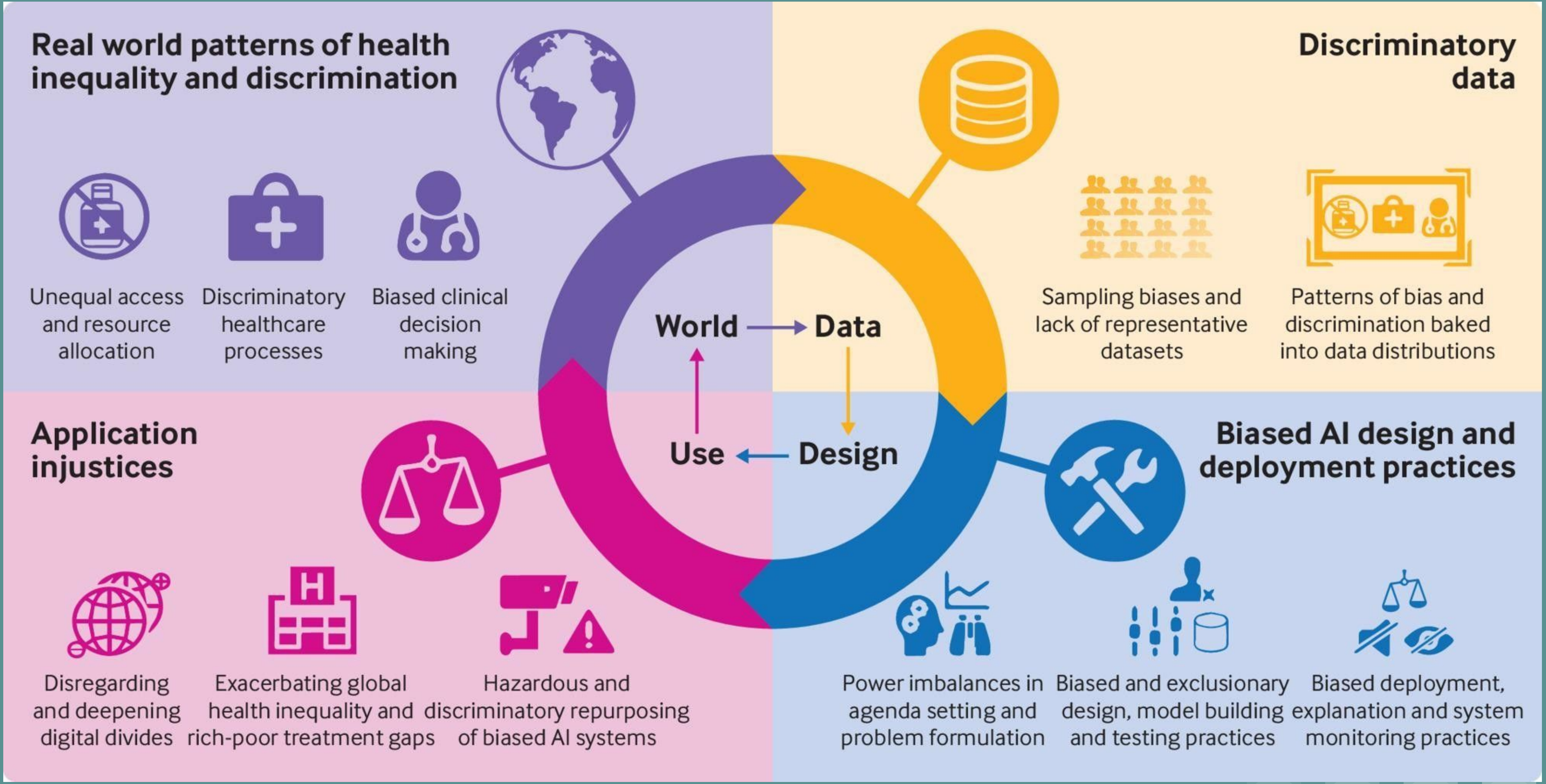
Mitigating Unwanted Biases With Adversarial Learning

Omkar R.D (2020A3PS0420P)
Pranav Dangi (2019A3PS0210P)
Vinayak Singh(2019B4A70606P)

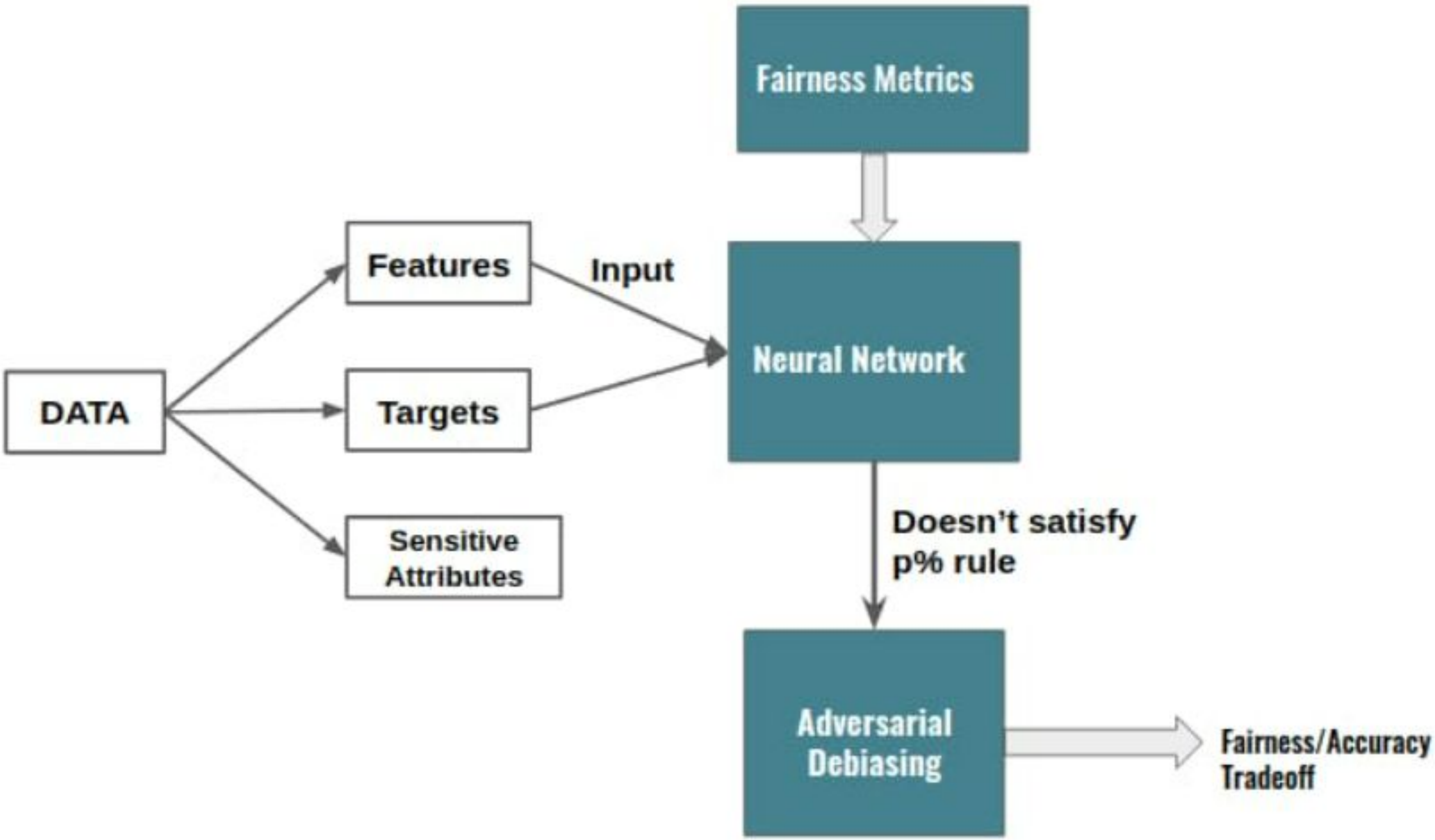

Real world patterns of health inequality and discrimination
Unequal access and resource allocation
Discriminatory healthcare processes
Biased clinical decision making
Discriminatory data
Sampling biases and lack of representative datasets
Patterns of bias and discrimination baked into data distributions
World
Data
Use
Design
Application injustices
Disregarding and deepening digital divides
Exacerbating global health inequality and rich-poor treatment gaps
Hazardous and discriminatory repurposing of biased AI systems
Biased AI design and deployment practices
Power imbalances in agenda setting and problem formulation
Biased and exclusionary design, model building and testing practices
Biased deployment, explanation and system monitoring practices

About the Paper

The objective is to maximize the predictors ability to predict Y while minimizing the adversary's ability to predict Z, a concept more commonly known as Adversarial Debiasing

Fairness:

Fairness in machine learning refers to the various attempts at correcting algorithmic bias based on machine learning models. Decisions made by computers after a machine-learning process may be considered unfair if they were based on variables considered sensitive.


Fairness Metrics
Features
Input
Neural Network
DATA
Targets
Sensitive Attributes
Doesn't satisfy p% rule
Adversarial Debiasing
Fairness/Accuracy Tradeoff

Motivation

Adverserial Learning:

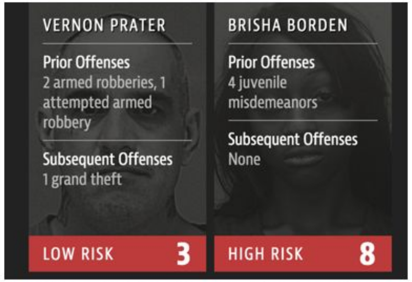
1. Adversarial AI Works with Minimal Labeled Data Pools
2. training with Less Human Supervision
3. High Fidelity Results

Why we chose this paper

1. The opportunity to learn and implement the concept of GANs

Why we chose this paper

2. The impact to deep rooted societal problems and beyond



Definitions of Fairness

Parity

$$P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z)$$

Odds Fairness

$$P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$$

Table 3. Lilliputian applicants (90% are qualified)

	Qualified	Unqualified
Admitted	45	2
Rejected	45	8
Total	90	10

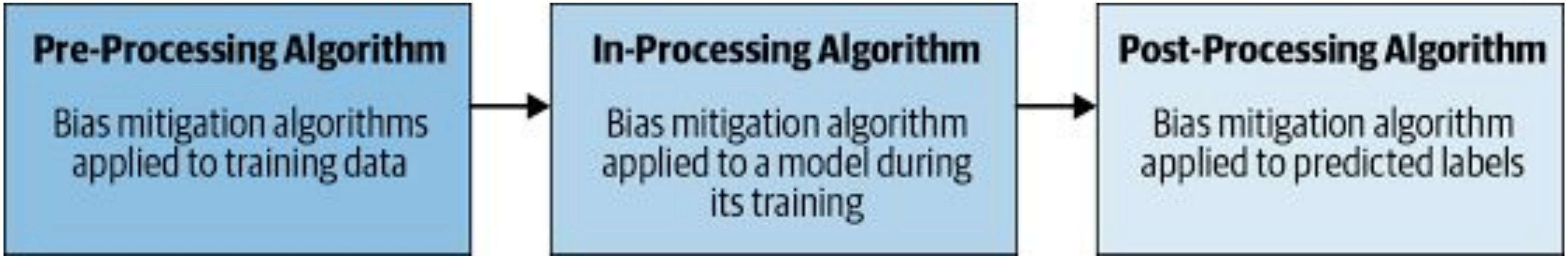
Percentage of qualified students admitted: 45/90 = 50%
Percentage of unqualified students rejected: 8/10 = 80%
Total percentage of Lilliputian students admitted: (45+2)/100 = 47%

Table 4. Brobdingnagian applicants (10% are qualified):

	Qualified	Unqualified
Admitted	5	18
Rejected	5	72
Total	10	90

Percentage of qualified students admitted: 5/10 = 50%
Percentage of unqualified students rejected: 72/90 = 80%
Total percentage of Brobdingnagian students admitted: (5+18)/100 = 23%

Methods of Removing Bias



Adversarial Debiasing:

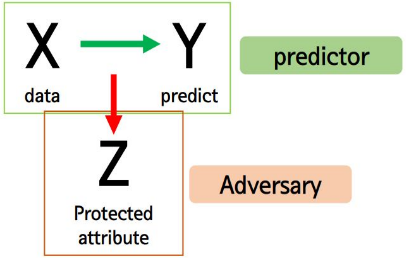
It tries to predict, based upon the predictions of your first model, the sensitive attribute. Ideally, in a situation without bias, this adversarial model should not be able to predict well the sensitive attribute. The adversarial model, therefore, guides modifications of the original model (via parameters and weighting) that **weakens the predictive power** of the adversarial model **until it cannot predict the protected attributes well based upon the outcomes.**

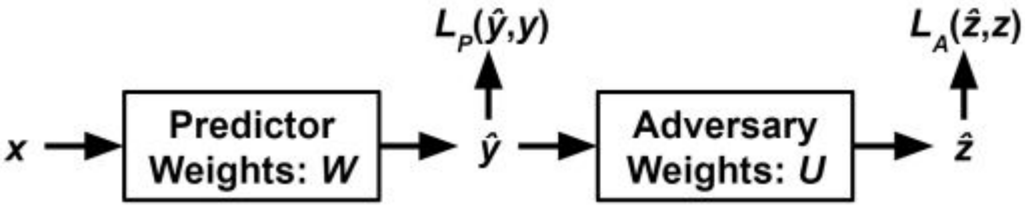
Advantages:

1. The first advantage of this method is that you directly intervene at the learning stage of the modeling workflow. In addition, it can be applied to both classification and regression.

2. The second advantage is that this is this approach is applicable to different fairness definitions as well.

Adversarial Network Architecture





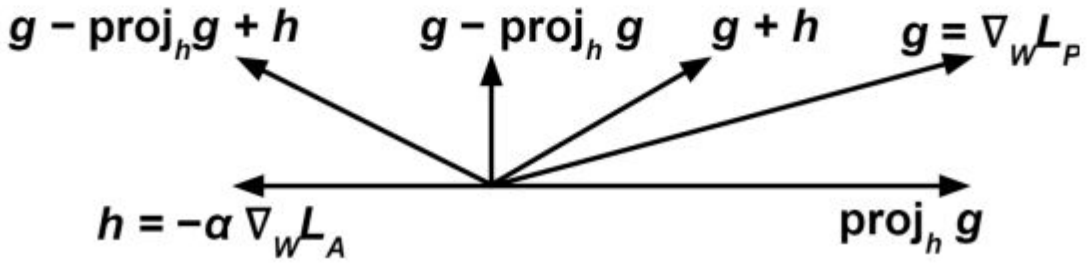
- Weight Updating **U**

$$\nabla_U L_A$$

- Weight Updating **W**

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A$$

Intuition of the Learning Algorithm



Without the projection term, the predictor would move in the direction of **g+h,** which actually helps the adversary.

Change with the projection term: Predictor never moves in direction that helps adversary

Specific Properties of the Model

- Generality
- Model Agnostic
- Optimality

Future Works

- Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction

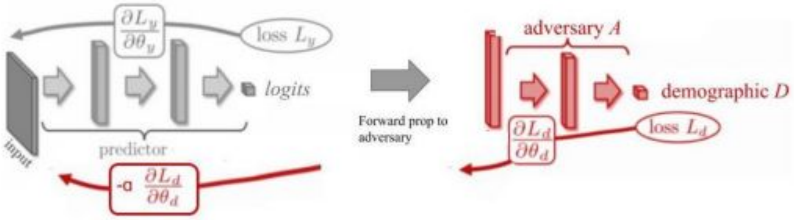


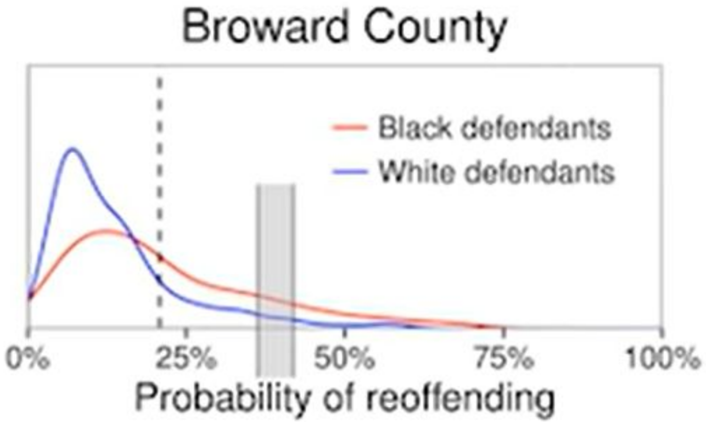
Figure 1. Diagram of our adversarial model structure.

MODEL	HIGH RISK GAP	FN GAP	FP GAP
COMPAS SCORES (OUR TEST SET)	0.18	0.22	0.17
OUR RECIDIVISM MODEL	0.21	0.27	0.15
OUR CHOSEN ADVERSARIAL MODEL	**0.02**	**0.02**	**0.01**

Understanding COMPAS Dataset

- Used to predict Likelihood of recidivism among criminal offenders

- Criticism for holding racial bias



Theoretical Guarantees

Proposition 1. *Let the predictor, the adversary, and their weights* W, U *be defined according to Section 3 Let* $L_A(W, U)$ *be the adversary's loss, convex in* U, *concave in* W,[4] *and continuously differentiable everywhere.*

Then, $L_A(W^*, U^*) = L_A(W^*, U_0)$. *That is, the adversary gains no advantage from using the weights for* $\hat{Y}$.

Theoretical Guarantees

Proposition 2:

Perfect Demographic Parity: Adversary achieves loss H(Z), the entropy of Z.

Proposition 3:

Perfect Equality of Odds:

Our Implementation

Dataset used: COMPAS, Original Paper Implementation: UCIAdult

Model Architecture:

1. Classifier Model: 200 ReLU Units (100x2)
2. Adversary Model: 100 ReLU Nodes
3. Fitting
4. Predicting: Classifier Accuracy & Fairness Metrics

Classifier Model:

- Optimized bias and weights initialization using GlorotUniform
- Compute the classifier predictions for the outcome variable.
- return pred_label, pred_logit

Adversary Model:

- Compute the adversary predictions for the protected attribute based on fairness_def = “parity” OR “equal_odds”
- return pred_protected_attribute_label, pred_protected_attribute_logit

Results

1. Performance of Classifier

PERFORMANCE:

	precision	recall	f1-score	support
0	0.63	0.83	0.71	672
1	0.67	0.42	0.51	562
accuracy			0.64	1234
macro avg	0.65	0.62	0.61	1234
weighted avg	0.65	0.64	0.62	1234

Results

2. Fairness Metric

```
proportion of White people predicted to reoffend: 0.24087591240875914
proportion of Nonwhite people predicted to reoffend: 0.30741190765492105
        RATE GAP = -0.06653599524616191

TPR for White people: 0.34782608695652173
TPR for Nonwhite people: 0.4463840399002494
        TPR GAP = -0.09855795294372766

FPR for White people: 0.172
FPR for Nonwhite people: 0.17535545023696683
        FPR GAP = -0.003355450236966845
```

Results

2. Fairness Metric

```
BIAS:

Correlation between age and predicted label: -0.13488246652161495

Correlation between age and predicted label, conditional on true label=1: -0.08786945959957115

Correlation between age and predicted label, conditional on true label=0: -0.10181125409199346
```

Results

Tables of RATE GAP with and without debiasing

Without Debiasing	With Parity Fairness	With Equal Odds Fairness
-0.147	-0.063	-0.013
-0.174	-0.088	-0.003
-0.154	-0.106	-0.039

Results

Tables of TPR GAP with and without debiasing

Without Debiasing	With Parity Fairness	With Equal Odds Fairness
-0.204	-0.087	-0.025
-0.227	-0.134	-0.038
-0.203	-0.152	-0.063

Results

Tables of FPR GAP with and without debiasing

Without Debiasing	With Parity Fairness	With Equal Odds Fairness
-0.062	-0.002	0.034
-0.088	-0.038	0.057
-0.075	-0.023	0.014

Results

Correlation GAP (age) for Parity Fairness

Without Debiasing	With Parity Fairness	With Equal Odds Fairness
0.229	0.108	0.017
0.243	0.212	0.020
0.208	0.01	0.111

Code Implementation

Thank You !