An Auditor's Guide for Discrimination Detection

Federica Picogna

- Development of a decision-making workflow to address the problem of discrimination in AI outcomes

- Despite the progress made, identification of new challenges regarding the audit risk and the lack of a tolerance threshold. Undoubtedly, there may be numerous other issues to consider

Limperg Instituut

The Growing Impact of Artificial Intelligence: Benefits and Risks Tied with Its Use

Artificial Intelligence is affecting our decisions and our lifestyle every day

→ Benefits

Increased decision-making speed

Automated repetitive tasks

Boosted productivity

→ Downsides

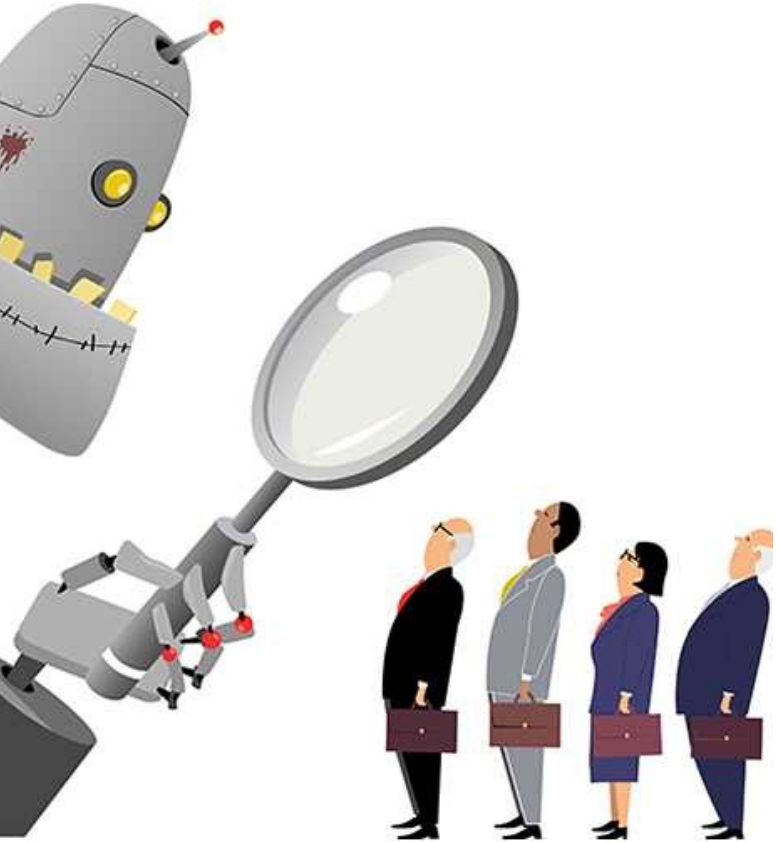
Privileged group vs Protected group

Possible discrimination

VU

Limperg Instituut

Understanding the Benefits and Risks of AI Use with a Real-Life Case Study: The Amazon Recruiting Tool



Can we expedite the hiring process?

Automatic selection of the candidates whose resumes contain the requirements for a certain job position

HOWEVER

AI brings to a systematic undervaluation of women's resumes for technical job

Balancing the Benefits and Risks of AI Use through Regulations like the AI Act

The Artificial Intelligence Act

→ Benefits of AI use

→ Respect for the rights recognized for all EU citizens

Who can verify the existence of this delicate balance? **The auditor**

What Are the Auditor's Challenges in AI Assessment? Fairness Definition and Statistical Tool Selection

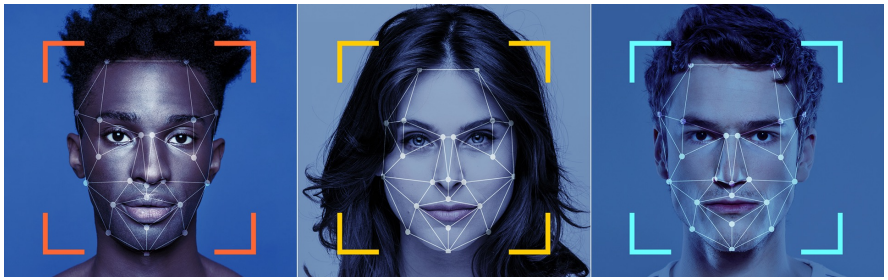
1. What is meant by "Fair"?

2. Which instrument should they use?

Limperg Instituut

The Ambiguity of the AI Act in Defining the Auditor's Task: What Does “Fairness” Mean?

> ***Diversity, non-discrimination and fairness means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law.***



The Use of Facial Recognition Technology in the Criminal Justice System:

- It Is Gender-Sensitive
- Should We More Closely Monitor Those Primarily Responsible for Crimes?

What Are the Auditor's Challenges in AI Assessment? Fairness Definition and Statistical Tool Selection

1. What is meant by "Fair"?

2. Which instrument should they use?

MEANING: Which fairness measure should they choose?

Understanding Fairness Measures through a Real-Case Example: Description of the COMPAS Dataset

- 6172 offenders
- 14 different information are included, covering details about the type of crime and information about the offenders
- Race: Sensitive Attribute
  - 2103 Caucasians
  - 4069 Non-Caucasians

Low risk of committing another crime	High risk of committing another crime
1281 Caucasian	822 Caucasian
2082 Non-Caucasian	1987 Non-Caucasian
3363 People in Total	2809 People in Total

The Majority of Fairness Measures Stem from the Confusion Matrix: Understanding it through the COMPAS Dataset

		Actual Value: n	Actual Value: p	
Predicted Value	$\hat{n}$	587 **True Negative (TN)**	245 **False Negative (FN)**	832 **Predicted Negative**
	$\hat{p}$	244 **False Positive (FP)**	467 **True Positive (TP)**	711 **Predicted Positive**
		831 **Actual Negative**	712 **Actual Positive**	1543 **Total**

Summary of the Quantities in the Confusion Matrix

- Negative = Label provided by the algorithm to indicate an offender who has not been arrested within 2 years of release.
- Positive = Label provided by the algorithm to indicate an offender who has been arrested within 2 years of release.
- True Negative (TN) = The number of offenders that the algorithm correctly predicts will be not rearrested within two years of release
- True Positive (TP) = The number of offenders that the algorithm correctly predicts will be rearrested within two years of release
- False Negative (FN) = The number of offenders that the algorithm incorrectly predicts will be not rearrested within two years of release but instead get rearrested
- False Positive (FP) = The number of offenders that the algorithm incorrectly predicts will be rearrested within two years of release but instead do not get rearrested

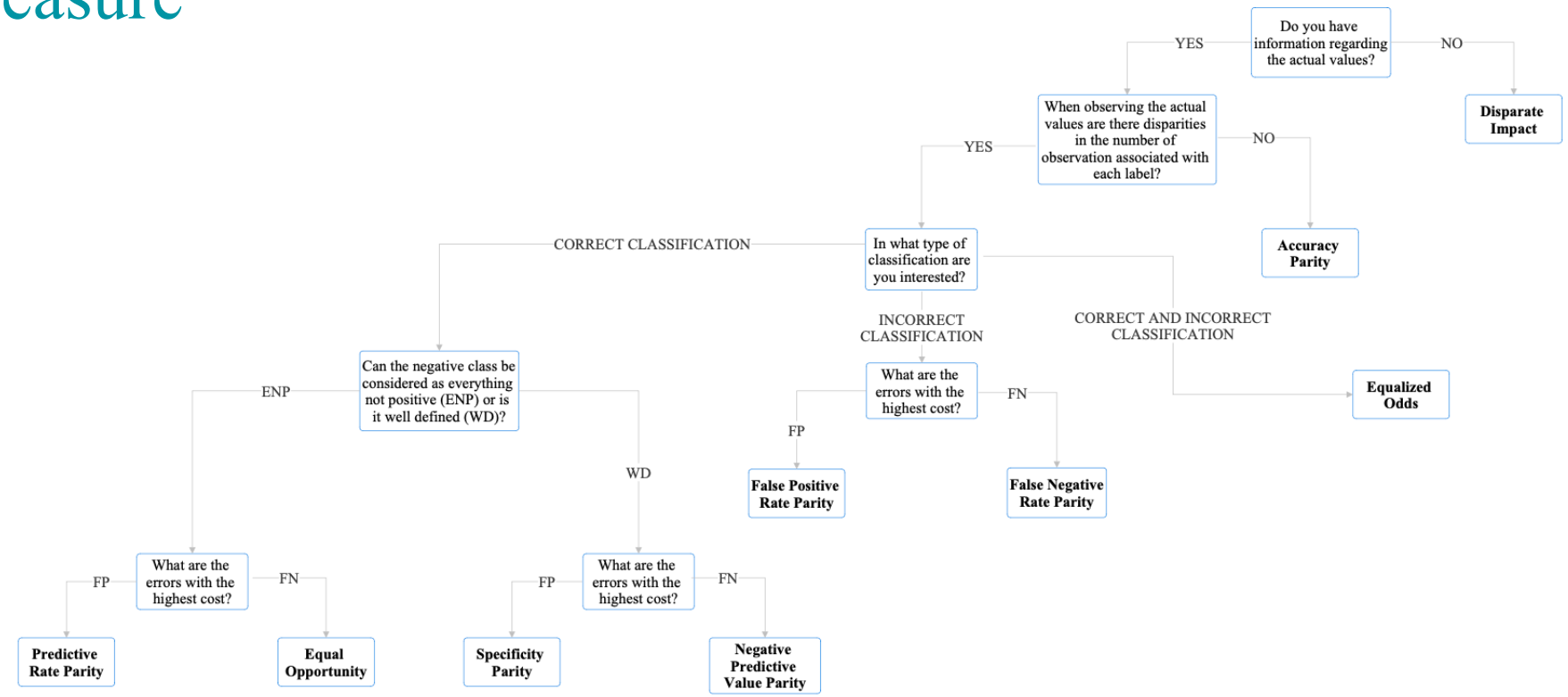
So many Fairness Measures Can Be Obtained by Combining the Quantities of the Confusion Matrix: Which One Should the Auditor Use?

"Column" based measures	TPR or Recall	$\frac{TP}{TP + FN}$
	FPR	$\frac{FP}{TN + FP}$
	TNR or Specificity	$\frac{TN}{TN + FP}$
	FNR	$\frac{FN}{TP + FN}$

"Row" based measures	Precision	$\frac{TP}{TP + FP}$
	Negative Predictive Value	$\frac{TN}{TN + FN}$
	Positive Rate	$\frac{TP + FP}{TP + FN + TN + FP}$

"Combined" measure	Accuracy	$\frac{TP + TN}{TP + FN + TN + FP}$

The Decision-Making Workflow the Auditor Can Use as a Guide in Selecting the Most Appropriate Fairness Measure



Limperg Instituut

Analyzing Discrimination in AI Predictions: A Practical Approach with the COMPAS Dataset

Privileged group – Caucasians

		Actual Value		
		n	p	
Predicted Value	$\hat{n}$	248	95	343
	$\hat{p}$	61	105	166
		309	200	509

Protected group – Not Caucasians

		Actual Value		
		n	p	
Predicted Value	$\hat{n}$	349	150	499
	$\hat{p}$	173	362	535
		522	512	1034

How different are the AI's predictions for Caucasian versus non-Caucasian offenders?

- A confusion matrix is obtained for each group: how AI correctly and incorrectly classifies the offenders
- Auditor's task: answering decision-making workflow questions to obtain the best fairness measure to use and use it to draw conclusions on potential discrimination in AI outcomes.

Limperg Instituut

Do We Have Discrimination in the COMPAS Dataset? Application of False Positive Rate Parity as a Fairness Measure

False Positive Rate for Caucasians :

$$\frac{FP}{TN + FP} = \frac{61}{61 + 248} = 0.197$$

The number of offenders that AI incorrectly predicts will be rearrested within two years of release but instead do not get rearrested.

False Positive Rate for Not Caucasians:

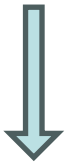
$$\frac{FP}{TN + FP} = \frac{173}{173 + 349} = 0.331$$

The total number of offender that were NOT rearrested within two years of release given by the number of offender that AI correctly predicts will be not rearrested within two years of release and the number of offenders that AI incorrectly predicts will be rearrested within two years of release but instead do not get rearrested

Addressing Additional AI Challenges for Future Development: The Need for a Tolerance Threshold for Tolerable Unfairness

1. When does the difference can be interpreted as evidence of discrimination in the AI outcome?

2. How can the auditor compare different metrics?



Tolerable Difference
Audit Risk

Disparate Impact: The Only Fairness Measure with a Threshold That However Lacks a Statistical Interpretation

Positive Rate for Caucasians :

$$\frac{TP + FP}{TP + TN + FP + FN} = \frac{166}{509} = 0.326$$

Positive Rate for Not Caucasians:

$$\frac{TP + FP}{TP + TN + FP + FN} = \frac{535}{1034} = 0.517$$

Disparate Impact:

$$\frac{PRC}{PRNC} = \frac{0.326}{0.517} = 0.630 < 0.80$$

For each group, the total number of offenders that AI correctly and incorrectly predicts will be rearrested within two years of release divided by the total amount of offenders

MEANING: How many offenders does AI think will be rearrested among all those for whom information is available? And how does this perception vary between the two groups? Does AI believe that Caucasian offenders will behave better?

WHAT WE DID SO FAR:

- Development of a decision-making workflow for the auditor to select the appropriate measure for evaluating discrimination

WHAT WE WILL WORK ON:

- Translation of the concept of Audit Risk into the algorithmic AI field
- Development of a Tolerance Difference in the algorithmic AI field

Thank you for the attention!