Towards Visualising Procedural Fairness in Automated Decision Making

Vis-Empowered Human-in-the-Loop AI Focus Period

Vladimiro González-Zelaya
30 April 2025
Linköping University

Flavours of Fairness

Distributive Fairness

- Focus on outcomes
- Easy to quantify
- Lots of literature

Procedural Fairness

- Focus on process
- Hard to quantify
- Just a few papers

Flavours of Fairness

Distributive Fairness

- Focus on outcomes
- Easy to quantify
- Lots of literature

Procedural Fairness

- Focus on process
- Hard to quantify
- Just a few papers

Flavours of Fairness

Distributive Fairness

- Focus on outcomes
- Easy to quantify
- Lots of literature

Procedural Fairness

- Focus on process
- Hard to quantify
- Just a few papers

Distributive Fairness

Protected Attribute (PA)

Newcastle University

Definition

The **Protected Attributes (PAs)** of a dataset are the features prone to an unjustified discriminatory decision

Examples

- Race or Skin Colour
- Sex or Gender
- Age
- Income Level
- Education

Some Fairness Definitions

Newcastle University

Fairness Through Unawareness

$$\hat{Y} = f\left(X \setminus \{\boldsymbol{PA}\}\right)$$

Demographic Parity

$$P(\hat{Y} = 1 \mid \boldsymbol{PA} = 0) = P(\hat{Y} = 1 \mid \boldsymbol{PA} = 1)$$

Equality of Opportunity

$$P(\hat{Y} = 1 \mid \boldsymbol{PA} = 0,\ Y = 1) = P(\hat{Y} = 1 \mid \boldsymbol{PA} = 1,\ Y = 1)$$

Fairness definitions are usually **incompatible** with each other

Some Fairness Definitions

Fairness Through Unawareness

The PA is not explicitly used during the decision process

Demographic Parity

Same **positive rate** across PA groups (*equality of outcomes*)

Equality of Opportunity

Same **true positive rate** across PA groups

Fairness definitions are usually **incompatible** with each other

Some Fairness Definitions

Newcastle University

Fairness Through Unawareness

The PA is not explicitly used during the decision process

Demographic Parity

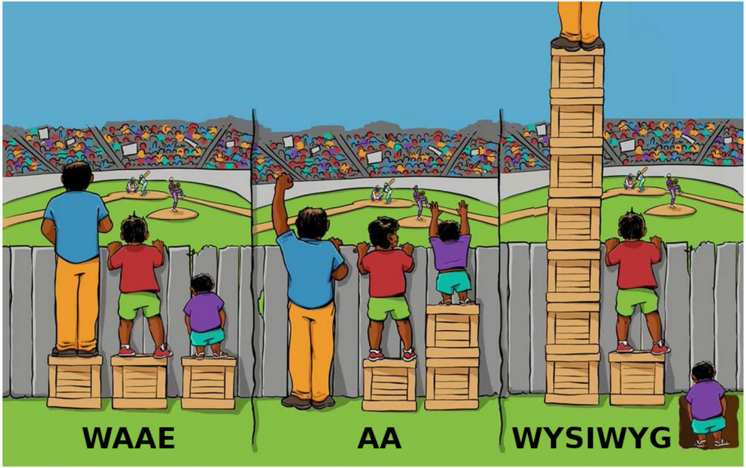
Same **positive rate** across PA groups (*equality of outcomes*)

Equality of Opportunity

Same **true positive rate** across PA groups

Fairness definitions are usually **incompatible** with each other

World View Alignment



The Evolution of an Accidental Meme — `https://link.medium.com/eFYERDAJNU`

Ways to Correct Unfairness

Pre-Processing Modify the training set to “sample from a better world”

In-Processing Add *constraints* or *regularisation terms* to improve fairness

Post-Processing Adjust the predictions after fitting the model

Ways to Correct Unfairness

Pre-Processing Modify the training set to “sample from a better world”

In-Processing Add *constraints* or *regularisation terms* to improve fairness

Post-Processing Adjust the predictions after fitting the model

Population Subgroups

Protected Attribute

- Favoured
- Unfavoured

Class

- Positive
- Negative

$$U+ \quad F+$$

$$U- \quad F-$$

Population Subgroups

Protected Attribute

- Favoured
- Unfavoured

Class

- Positive
- Negative

$$U+ \quad F+$$

$$U- \quad F-$$

Population Subgroups

Protected Attribute

- Favoured
- Unfavoured

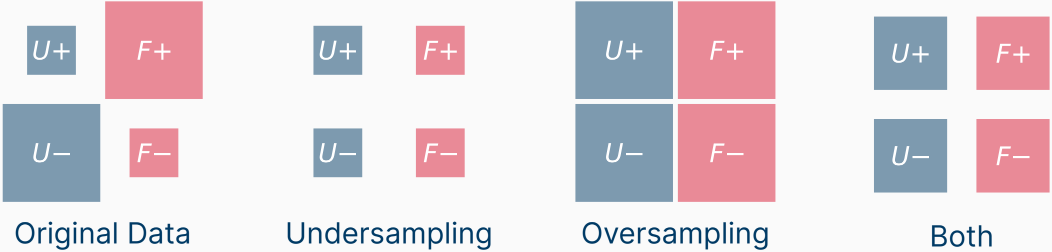
Class

- Positive
- Negative

$$U+ \quad F+$$

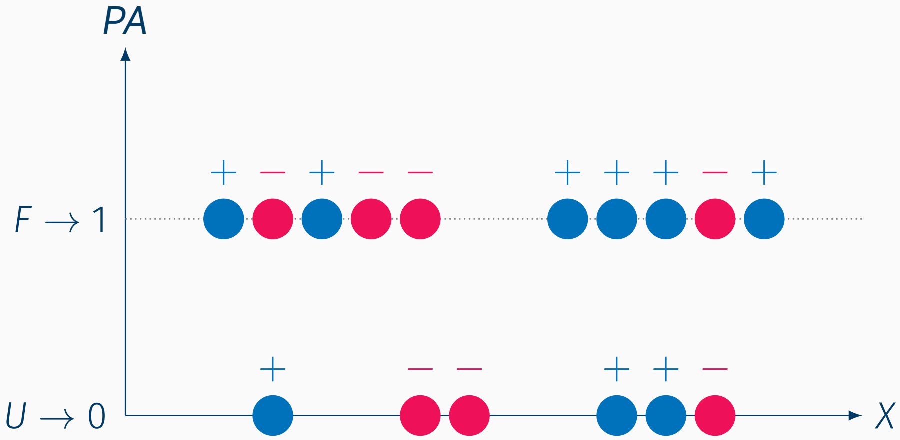
$$U- \quad F-$$

Fairness Correction Through Resampling[1]

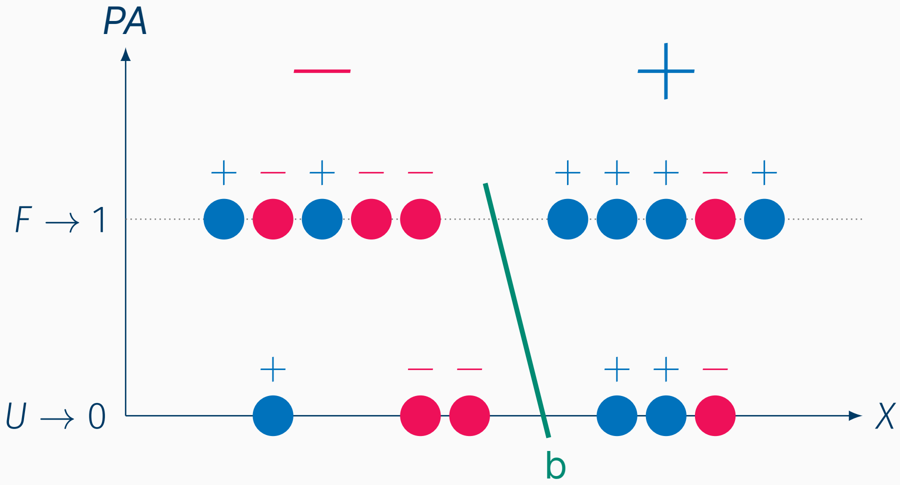


[1] Vladimiro González-Zelaya, Julián Salas, Dennis Prangle, and Paolo Missier (2021). "Optimising Fairness through Parametrised Data Sampling". In: *24th International Conference on Extending Database Technology, EDBT 2021*.

An Optimal Linear Classifier (Accuracy)



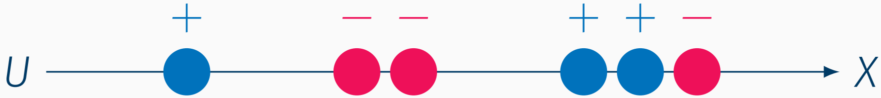
An Optimal Linear Classifier (Accuracy)



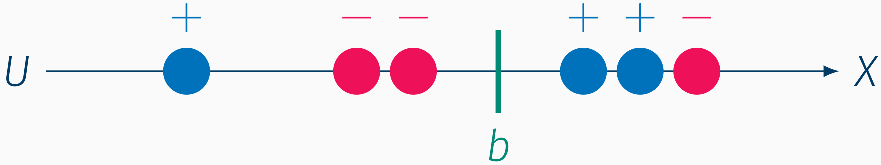
The positive predictions for U increase by:

- Undersampling negative instances
- Oversampling positive instances

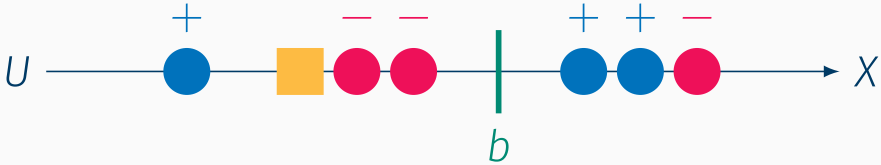
Predictions for U



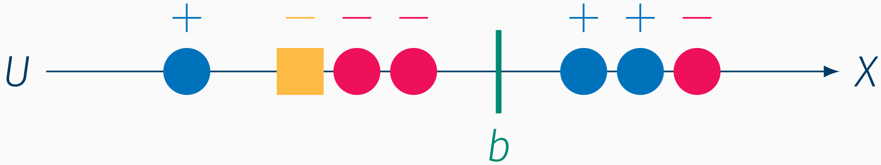
Predictions for U



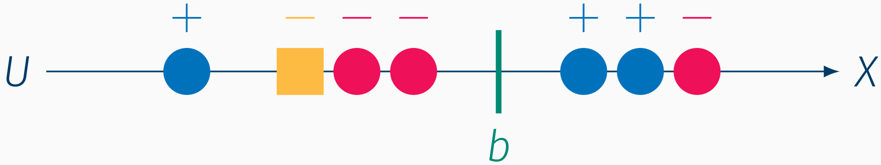
Predictions for *U*



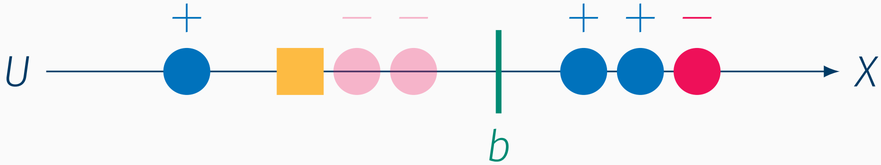
Predictions for U



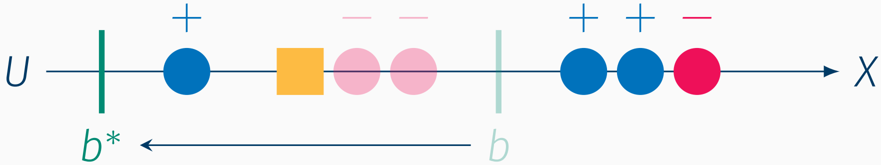
Negative Undersampling Left of b



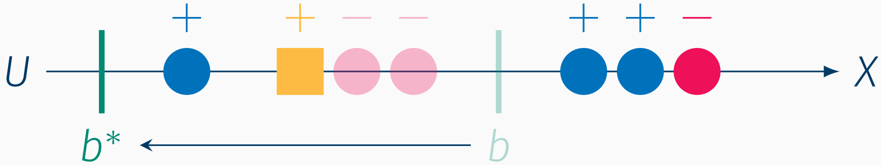
Negative Undersampling Left of b



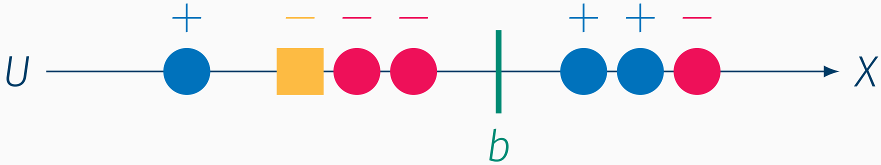
Negative Undersampling Left of b



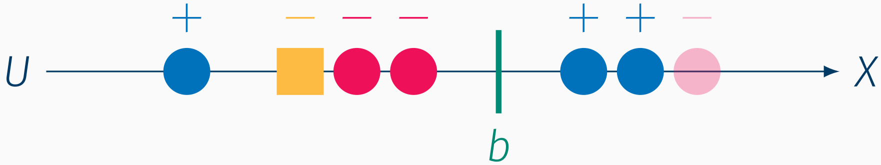
Negative Undersampling Left of b



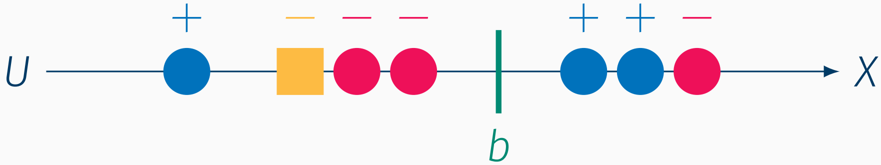
Negative Undersampling Right of *b*



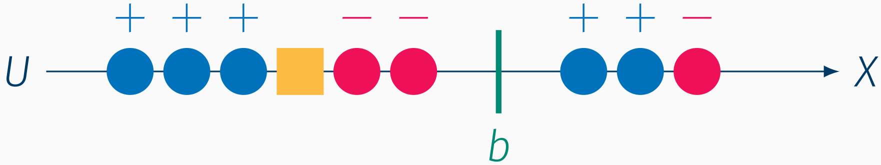
Negative Undersampling Right of b



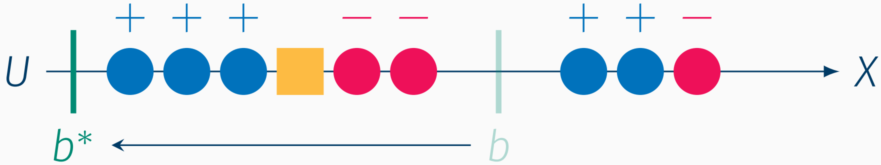
Positive Oversampling Left of b



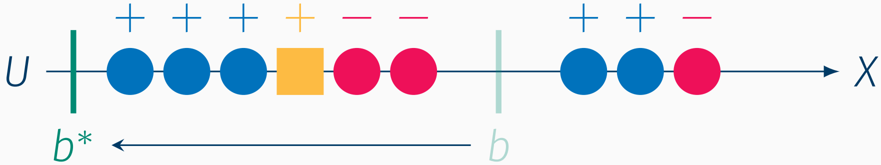
Positive Oversampling Left of b



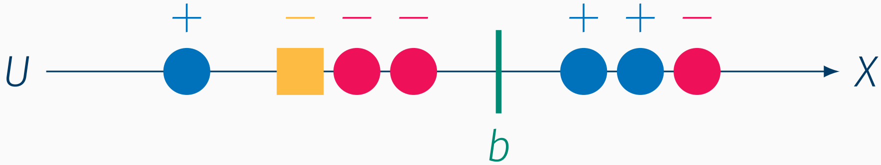
Positive Oversampling Left of b



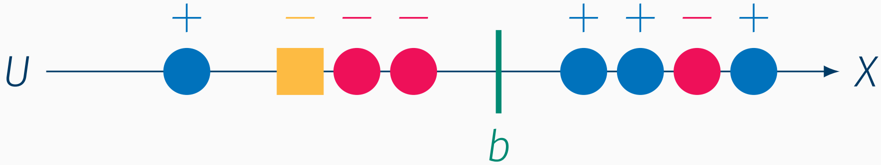
Positive Oversampling Left of b



Positive Oversampling Right of b



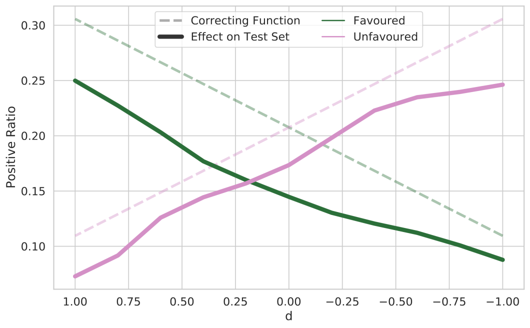
Positive Oversampling Right of b

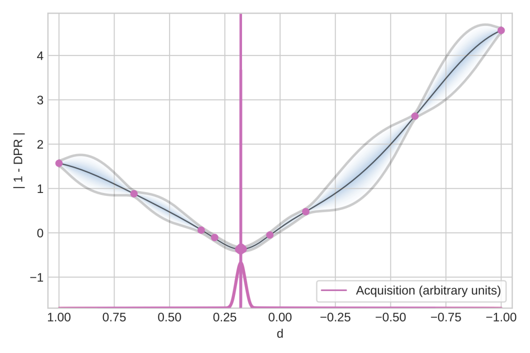


Analogously, positive predictions for F decrease by:

- Oversampling negative instances
- Undersampling positive instances

The Optimal Resampling Parameter





Dealing with Multiple PAs

	Age	Country	Gender	Race	Combined PA
	(20-30]	Mexico	Male	Latino	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR					
Data Set PR	0.3	0.3	0.3	0.3	
Difference					

Dealing with Multiple PAs

	Age	Country	Gender	Race	Combined PA
	(20-30]	Mexico	Male	Latino	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR					
Data Set PR	0.3	0.3	0.3	0.3	
Difference					

Dealing with Multiple PAs

Age	Country	Gender	Race	Combined PA
(20-30]	Mexico	Male	Latino	Unfavoured
(30-40]	Canada	Female	White	Favoured

Subgroup PR	0.2	0.3	0.4	0.1
Data Set PR	0.3	0.3	0.3	0.3
Difference				

Dealing with Multiple PAs

	Age	Country	Gender	Race	Combined PA
	(20-30]	Mexico	Male	Latino	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.2	0.3	0.4	0.1	
Data Set PR	0.3	0.3	0.3	0.3	
Difference					

Dealing with Multiple PAs

	Age	Country	Gender	Race	Combined PA
	(20-30]	Mexico	Male	Latino	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.2	0.3	0.4	0.1	
Data Set PR	0.3	0.3	0.3	0.3	
Difference	−0.1	+0.0	+0.1	−0.2	

Dealing with Multiple PAs

	Age	Country	Gender	Race	Combined PA
	(20-30]	Mexico	Male	Latino	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.2	0.3	0.4	0.1	
Data Set PR	0.3	0.3	0.3	0.3	
Difference	−0.1	+0.0	+0.1	−0.2	Sum = −0.2

Dealing with Multiple PAs

	Age	Country	Gender	Race	Combined PA
	(20-30]	Mexico	Male	Latino	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.2	0.3	0.4	0.1	
Data Set PR	0.3	0.3	0.3	0.3	
Difference	−0.1	+0.0	+0.1	−0.2	Sum = −0.2

Dealing with Multiple PAs

	Age	Country	Gender	Race	Combined PA
	(20-30]	Mexico	Male	Latino	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR					
Data Set PR	0.3	0.3	0.3	0.3	
Difference					

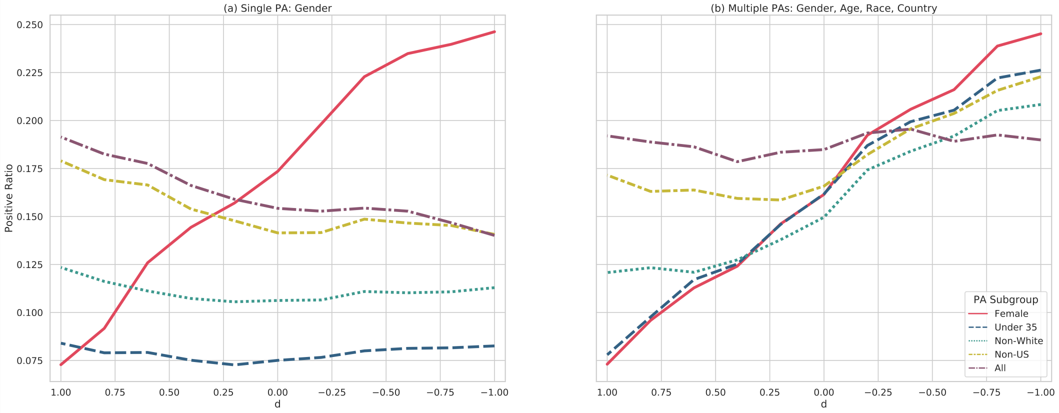
Dealing with Multiple PAs

	Age	Country	Gender	Race	Combined PA
	(20-30]	Mexico	Male	Latino	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.4	0.4	0.1	0.4	
Data Set PR	0.3	0.3	0.3	0.3	
Difference	$+0.1$	$+0.1$	-0.2	$+0.1$	Sum $= +0.1$

Dealing with Multiple PAs

	Age	Country	Gender	Race	Combined PA
	(20-30]	Mexico	Male	Latino	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.4	0.4	0.1	0.4	
Data Set PR	0.3	0.3	0.3	0.3	
Difference	+0.1	+0.1	−0.2	+0.1	Sum = +0.1

Multiple PA Correction



Procedural Fairness

Ways a Process Can Be Fair (Rawls 1971)

Perfect Procedural Fairness

If followed correctly, a fair outcome is guaranteed

Imperfect Procedural Fairness

If followed correctly, a fair outcome is likely

Pure Procedural Fairness

Fairness is given by the process itself, outcomes are irrelevant

Automated Decision Making (ADM) Perceptions

- Generally considered less fair than Human Decision Making (HDM) (Acikgoz, Davison, Compagnone, and Laske 2020)
- Perceived as more fair for mechanical tasks (Lee 2018)
- Can by more consistent and accurate (Dawes, Faust, and Meehl 1989)
- Has no emotions-related bias (Martínez-Miranda and Aldea 2005)

The Justice Rules (Leventhal 1980)

- 🎯 Accuracy
- ✅ Consistency
- 👨‍👩‍👧‍👦 Representativeness
- ⚖️ Bias Suppression
- 🛠️ Correctability
- 🧭 Ethicality

260 McNuggets? McDonald's Ends A.I. Drive-Through Tests Amid Errors

Ordering mistakes frustrated customers during nearly three years of tests. But competitors like White Castle and Wendy's say their A.I. ordering systems have been highly accurate.

Listen to this article · 6:53 min Learn more

Share full article

Damian Dovarganes/Associated Press

The Justice Rules (Leventhal 1980)

- 🎯 Accuracy
- ✅ Consistency
- 👨‍👩‍👧‍👦 Representativeness
- ⚖️ Bias Suppression
- 🛠️ Correctability
- 🧭 Ethicality

Can we predict when and where a crime will take place?

30 October 2018

Can algorithms really predict where new crimes will take place?

The Justice Rules (Leventhal 1980)

Newcastle University

- 🎯 Accuracy
- ✅ Consistency
- 👨‍👩‍👧‍👦 Representativeness
- ⚖️ Bias Suppression
- 🛠️ Correctability
- 🧭 Ethicality

Amazon scrapped 'sexist AI' tool

10 October 2018

The algorithm repeated bias towards men, reflected in the technology industry

The Justice Rules (Leventhal 1980)

- 🎯 Accuracy
- ✅ Consistency
- 👨‍👩‍👧‍👦 Representativeness
- ⚖️ Bias Suppression
- 🛠️ Correctability
- 🧭 Ethicality

A STAT INVESTIGATION

IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close

By Casey Ross and Ike Swetlitz Sept. 5, 2017

EROS DERVISHI FOR STAT

The Justice Rules (Leventhal 1980)

- 🎯 Accuracy
- ✅ Consistency
- 👨‍👩‍👧‍👦 Representativeness
- ⚖️ Bias Suppression
- 🛠️ Correctability
- 🧭 Ethicality

Robodebt: Illegal Australian welfare hunt drove people to despair

7 July 2023

The "Robodebt" policy vilified recipients of welfare, an inquiry has found

The Justice Rules (Leventhal 1980)

- 🎯 Accuracy
- ✅ Consistency
- 👨‍👩‍👧‍👦 Representativeness
- ⚖️ Bias Suppression
- 🛠️ Correctability
- 🧭 Ethicality

Clearview AI fined by Dutch agency for facial recognition database

By **Reuters**

September 3, 2024 9:21 PM GMT+1 · Updated 7 months ago

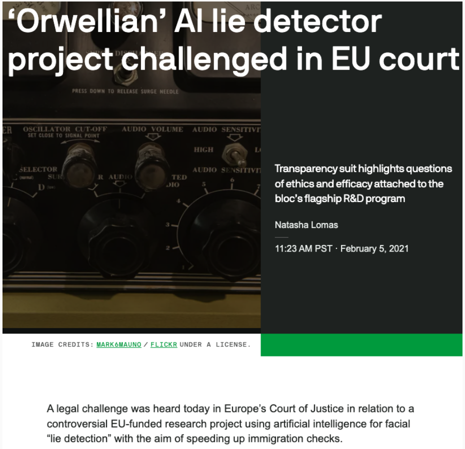
Aa

AI Artificial Intelligence words are seen in this illustration taken, May 4, 2023. REUTERS/Dado Ruvic/Illustration/File Photo Purchase Licensing Rights

The Extended Justice Rules

Newcastle University

- 🎯 Accuracy
- ✅ Consistency
- 👨‍👩‍👧‍👦 Representativeness
- ⚖️ Bias Suppression
- 🛠️ Correctability
- 🧭 Ethicality
- 🤔 Explainability
- 🪟 Transparency
- 🤝 Accountability


‘Orwellian’ AI lie detector project challenged in EU court

Transparency suit highlights questions of ethics and efficacy attached to the bloc’s flagship R&D program

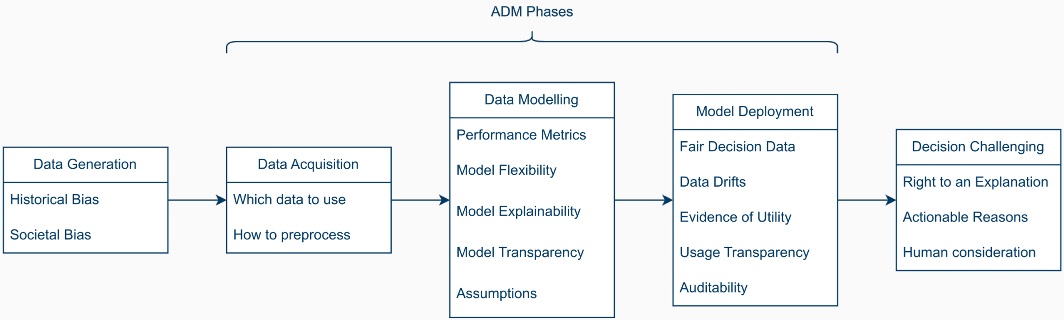
Natasha Lomas

11:23 AM PST · February 5, 2021

IMAGE CREDITS: MARK6MAUNO / FLICKR UNDER A LICENSE.

A legal challenge was heard today in Europe’s Court of Justice in relation to a controversial EU-funded research project using artificial intelligence for facial “lie detection” with the aim of speeding up immigration checks.

Justice Rules are Transversal to ADM Phases



Can the Rules Be Metricised (or Visualised)?

Newcastle University

(Kind of) Natural Metrics

Accuracy	Overall model performance
Consistency	Similar performance across groups
Representativity	Data vs population diversity
Correctability	Proportion of corrected decisions

Hard(er) to Metricise

Bias Supression	Funding sources, perceived bias
Ethicality	Adherence to legal frameworks
Transparency	Proportion of documented components
Explainability	Use of explainers, user comprehension

Can the Rules Be Metricised (or Visualised)?

Newcastle University

(Kind of) Natural Metrics

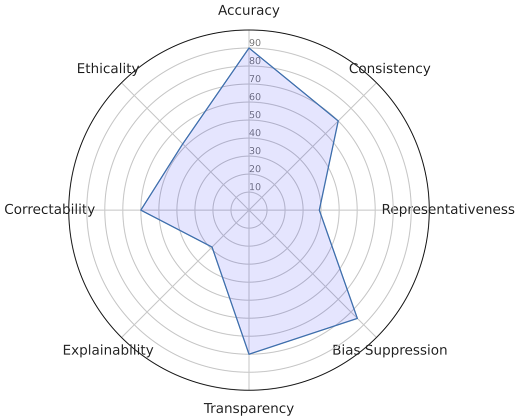
Accuracy	Overall model performance
Consistency	Similar performance across groups
Representativity	Data vs population diversity
Correctability	Proportion of corrected decisions

Hard(er) to Metricise

Bias Supression	Funding sources, perceived bias
Ethicality	Adherence to legal frameworks
Transparency	Proportion of documented components
Explainability	Use of explainers, user comprehension

Nutritional Labels / Procedural Radar

	Data Generation	Data Acquisition	Data Modelling	Model Deployment	Decision Challenging
Accuracy	90	95	90	85	90
Consistency	75	70	65	70	70
Representativeness	40	30	35	40	50
Bias Suppression	85	90	80	85	85
Transparency	85	80	75	80	80
Explainability	30	25	20	30	40
Correctability	50	60	65	60	65
Ethicality	60	50	45	50	55



What Needs to Be Done

Design Define metrics and visualisations

Test Run a fairness-perception study

Refine Identify and correct key concerns

Deploy Implement procedural dashboard

Conclusions

- Distributive fairness is easier to metricise
- However, must be aligned to world views
- Procedural complements distributive fairness for trust
- Starting point: Leventhal 1980’s rules for ADM
- Visualisations may bridge gap between concepts and perceptions

Thank You! Questions?

Conclusions

- Distributive fairness is easier to metricise
- However, must be aligned to world views
- Procedural complements distributive fairness for trust
- Starting point: Leventhal 1980's rules for ADM
- Visualisations may bridge gap between concepts and perceptions

Thank You! Questions?

Conclusions

Newcastle University

- Distributive fairness is easier to metricise
- However, must be aligned to world views
- Procedural complements distributive fairness for trust
- Starting point: Leventhal 1980's rules for ADM
- Visualisations may bridge gap between concepts and perceptions

Thank You! Questions?

Conclusions

- Distributive fairness is easier to metricise
- However, must be aligned to world views
- Procedural complements distributive fairness for trust
- Starting point: Leventhal 1980's rules for ADM
- Visualisations may bridge gap between concepts and perceptions

Thank You! Questions?

Conclusions

- Distributive fairness is easier to metricise
- However, must be aligned to world views
- Procedural complements distributive fairness for trust
- Starting point: Leventhal 1980's rules for ADM
- Visualisations may bridge gap between concepts and perceptions

Thank You! Questions?

Conclusions

- Distributive fairness is easier to metricise
- However, must be aligned to world views
- Procedural complements distributive fairness for trust
- Starting point: Leventhal 1980's rules for ADM
- Visualisations may bridge gap between concepts and perceptions

Thank You! Questions?