Algorithm Audit

Unsupervised bias scan tool

A quantitative method to inform qualitative bias testing

NGO Algorithm Audit

January 4th 2023

Overview of Algorithm Audit's bias scan tool

1. ## Problem description

- Problem 1 (quantitative) – Detecting higher-dimensional forms of differentiation
- Problem 2 (qualitative) – A persistent gap between general legal requirements and concrete AI practice

2. ## Solution

- Unsupervised bias scan tool to *detect* differentiation (quantitative)
- A deliberative approach to *establish* discrimination (qualitative)

3. ## Case study

- Disparities in a BERT-based Twitter disinformation classifier (quantitative)
- Audit commission: Assessing potentially unfair treatment by an AI classifier (qualitative)

4. ## Conclusion

+ contributors and endorsments

What is Algorithm Audit?

i) **Audit commissions** – Advising on ethical issues emerging in concrete algorithmic practices

ii) **Technical tooling** – Implement and tests technical tools to detect and mitigate bias in practice

iii) **Advocacy** – Sharing techno-legal knowledge with society and policy makers

Supported by:

EUROPEAN ARTIFICIAL INTELLIGENCE FUND

NL AI Coalition

SIDN fonds

Problem 1: The human mind is not equipped to detect higher-dimensional forms of algorithmic differentiation

The quantitative reasoning paradigm of AI…

Exploiting higher-dimensional correlations

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & \ddots & & \\ \vdots & & & \\ a_{n,1} & & & a_{n,n} \end{pmatrix}$$

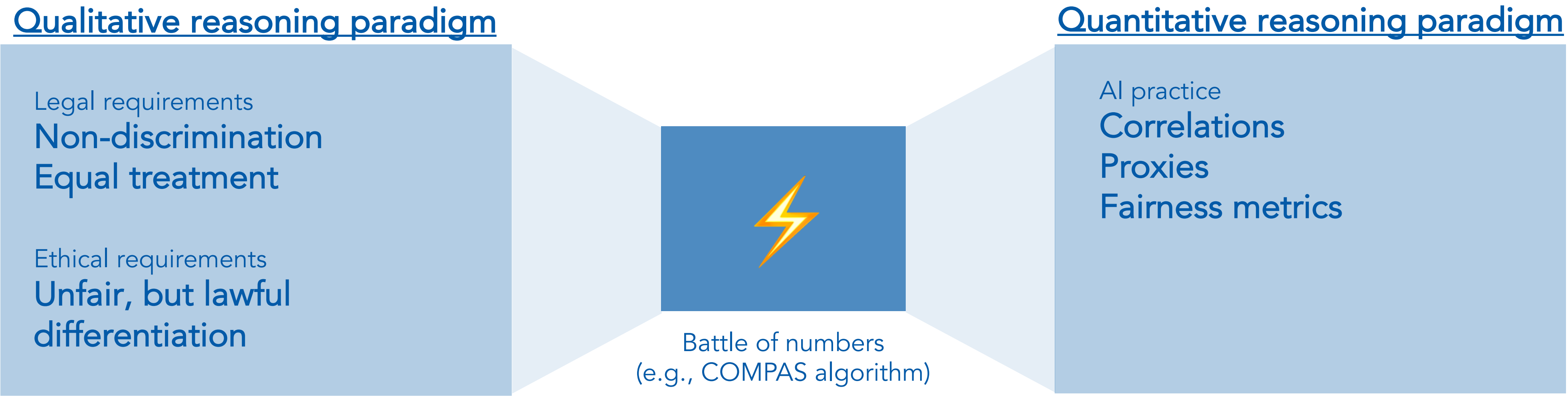
…poses challenges to assess fair treatment

1. How to detect disparities in the sheer data volume AI outputs?

2. How to detect differentiation upon new categories of people defined by a mixture of many data points (*ad hoc bias*)?

3. How to detect unfair differentiation when protected attributes are not available to compute group fairness metrics?

Problem 2: If differentiation is detected, a persistent gap remains between quantitative fairness metrics and qualitative interpretation



Qualitative reasoning paradigm

Legal requirements
Non-discrimination
Equal treatment

Ethical requirements
Unfair, but lawful differentiation

Battle of numbers
(e.g., COMPAS algorithm)

Quantitative reasoning paradigm

AI practice
Correlations
Proxies
Fairness metrics

● Independent audits of ‘gatekeeper’ platforms

● Conformity assessments

Urgently needed: normative guidelines to enforce upon new and existing legislation

🇪🇺

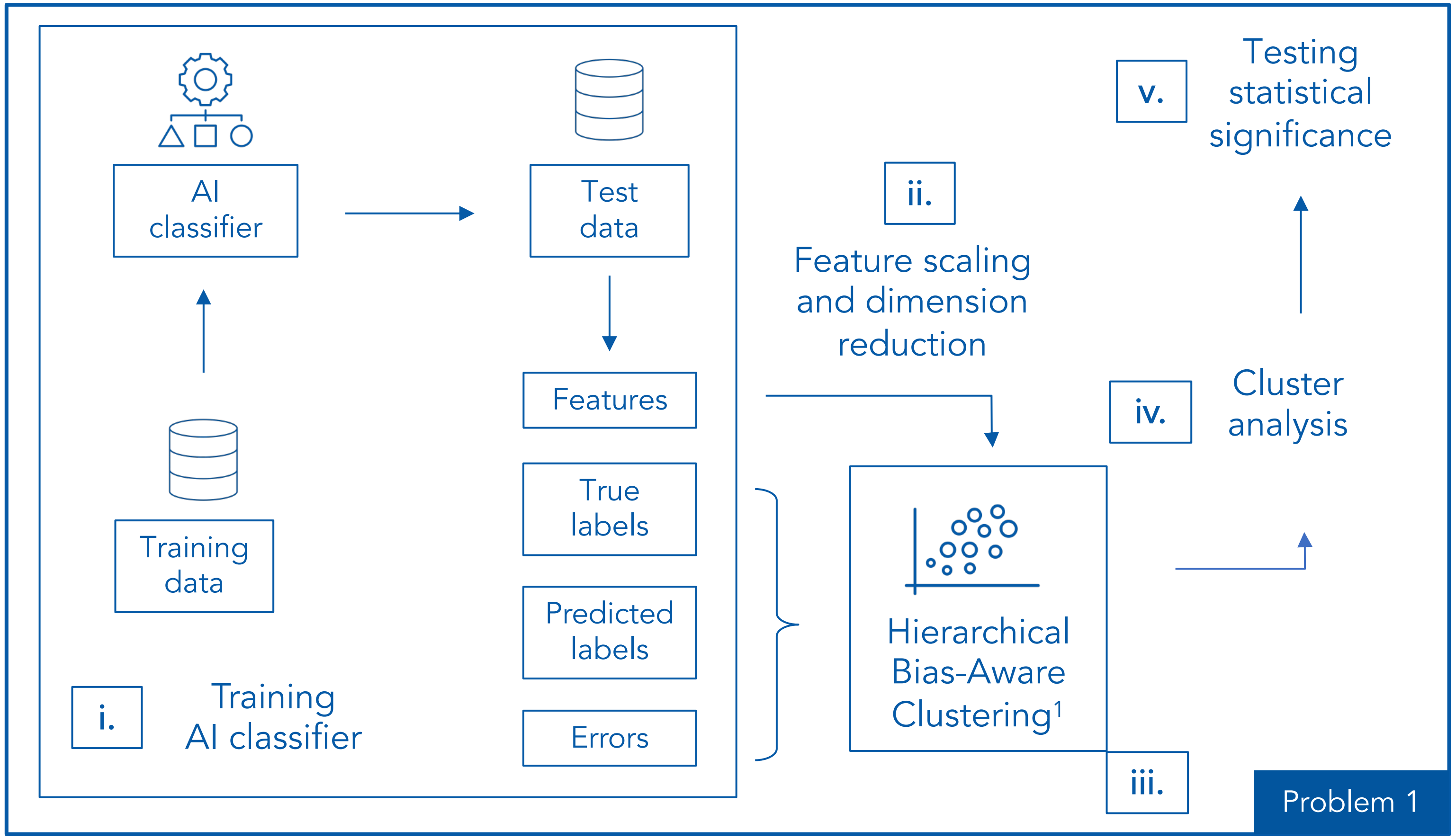
- LIVE Digital Services Act
- LIVE Digital Markets Act
- LIVE GDPR
- AI Act

🇺🇸

- NY Bias Law
- Equal Employment Opportunity LIVE
- State level laws LIVE

Solution: Quantitative method to *detect* differentiation (problem 1)
Qualitative approach to *establish* discrimination (problem 2)

Quantitative bias scan tool



i. Training AI classifier

Training data

AI classifier

Test data

Features

True labels

Predicted labels

Errors

ii. Feature scaling and dimension reduction

Hierarchical Bias-Aware Clustering[1]

iii.

iv. Cluster analysis

v. Testing statistical significance

Problem 1

Qualitative

A deliberative approach to establish algorithmic discrimination

1. **Identify issue**
   Identify potential discrimination by AI
2. **Audit commission**
   Form an independent and diverse commission of experts
3. **Analysis**
   Independent review of issue by audit commission
4. **Advice**
   Advice by audit commission is published and shared online

Problem 2

[1] Misztal, Indurkya, Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems, *Information Processing and Management* (2021)

Benefits of our quantitative-qualitative approach

Unsupervised machine learning → No data needed on protected characteristics of users.

Model agnostic → Works for all binary classification algorithms.

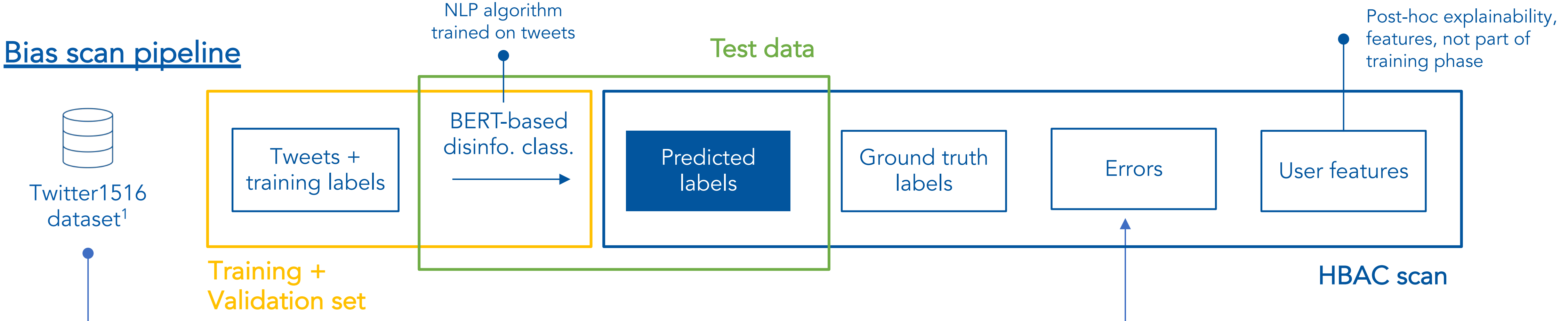
Bridging the quantitative and qualitative reasoning paradigm → Automated identification of potential bias allows human experts to assess observed disparities in a qualitative manner, subject to political, social and environmental traits.

Open-source and not for profit → Allowing the wider AI auditing ecosystem, e.g., data scientists, journalists, policy makers, public- and private auditors, to use quantitative methods to detect potential bias. Source code is available on GitHub.

The bias scan tool is available as a web application on our website. Creating instant a bias scan report for binary classifier data.

Detecting disparities on a self-trained BERT-based disinformation classifier, trained on the Twitter1516 dataset

Bias scan pipeline



NLP algorithm trained on tweets

Test data

Post-hoc explainability, features, not part of training phase

Twitter1516 dataset[1]

Tweets + training labels

BERT-based disinfo. class. →

Predicted labels

Ground truth labels

Errors

User features

Training + Validation set

HBAC scan

Twitter1516 dataset[1]

n=1.057 verified tweets (false or true)

5 content features:

- Length tweet
- #URLs
- #mentions
- #hashtags
- Sentiment score[2]

3 user features:

- Verified profile
- #followers
- User engagement[3]

Case study

False Positive Rate (FPR) bias metric
True content predicted to be false, proportional to all true content

Test 1: FPR metric

Bias = $\mathrm{FPR}_{\text{cluster}} - \mathrm{FPR}_{\text{rest of dataset}}$

False Negative Rate (FNR) bias metric
False content predicted to be true, proportional to all false content

Test 2: FNR metric

Bias = $\mathrm{FNR}_{\text{cluster}} - \mathrm{FNR}_{\text{rest of dataset}}$

[1] Liu, Xiaomo and Nourbakhsh, Armineh and Li, Quanzhi and Fang, Rui and Shah, Sameena, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (2015)

[2] Based on the VADER sentiment analysis tool, https://github.com/cjhutto/vaderSentiment

[3] Vosoughi, S., Roy, D., and Aral, S.: The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

Results: Disparities of a self-trained BERT-based Twitter disinformation classifier

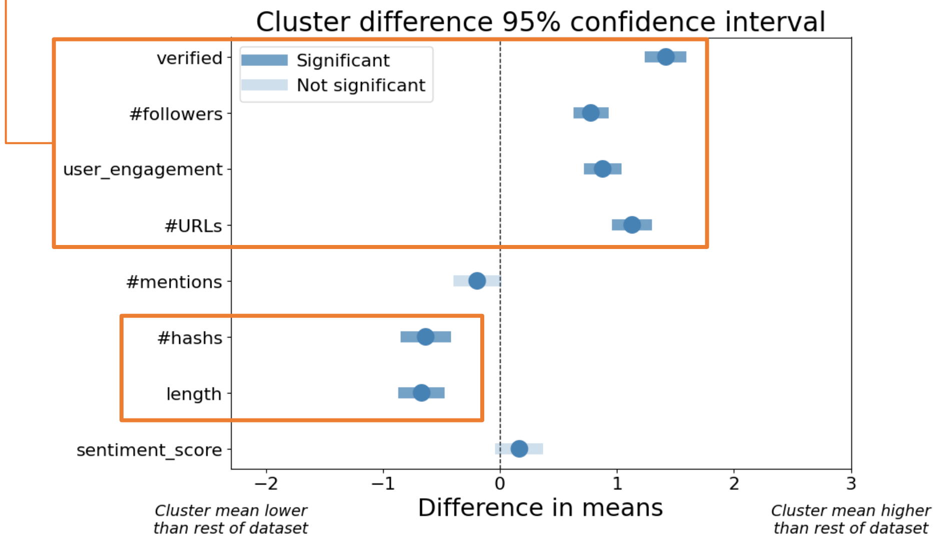
More details on [GitHub]

FPR scan

Cluster with highest rate of FPs: 0.08
#elements in highest biased cluster: 249

The cluster with the following features faces more FP classifications:

- Above average verified profiles, #followers, user engagement, #URLs
- Below average #hashags, tweet length

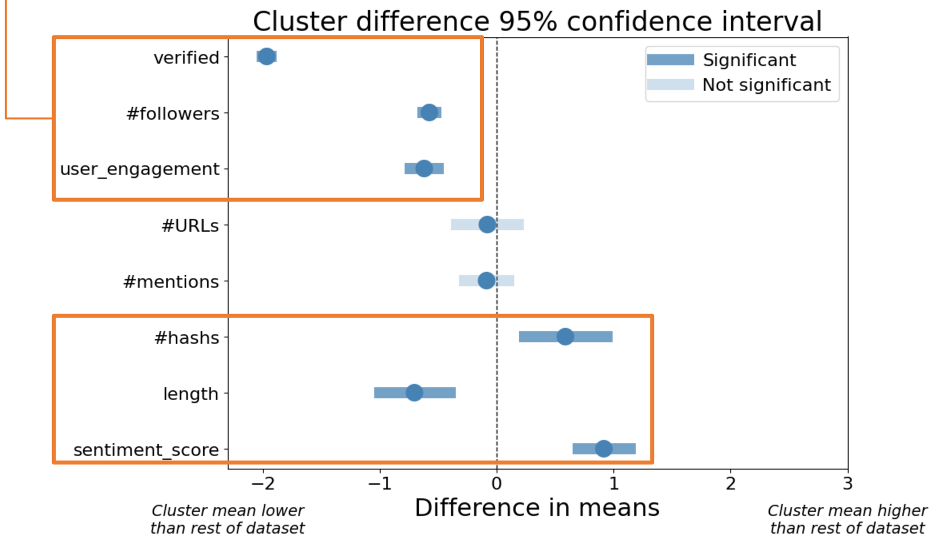


FNR scan

Cluster with highest rate of FNs: 0.13
#elements in highest biased cluster: 46

The cluster with the following features faces more FN classifications:

- Above average #hashtags, sentiment score;
- Below average verified profile, #followers, user engagement and tweet length



Audit commission: Qualitative assessment of potential unfair treatment by an AI classifier

Draft

Normative questions to establish unfair treatment

1. Is there an indication that one of the statistically significant features, or a combination of the features, is critically linked to one or multiple protected grounds?

2. Are False Positive classifications as harmful as False Negative classifications in this context?

3. Can the measured disparate treatment be justified given the aim pursued?

4. Considering the disparate treatment of users with a verified profile, above average sentiment score and/or below average number of URLs used in their tweets, could the observed disparate treatment be perceived as ethically undesirable?

Audit commission

Expert A

Expert B

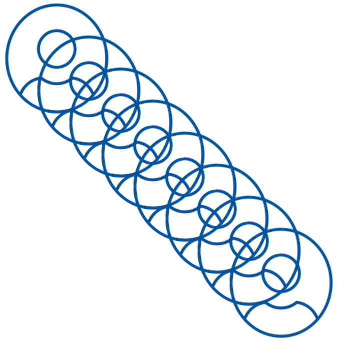
Expert C

Expert D

Conclusion: To be included once available

Audit commissions convenes
in Jan-Feb 2023, to elaborate on the
questions formulated in slide 9.

Contributors and endorsments



Algorithm Audit

Want to know more?
Get involved
Contact us!

info@algorithmaudit.eu
www.algorithmaudit.eu

https://www.linkedin.com/company/algorithm-audit/

https://github.com/NGO-Algorithm-Audit