Exposing the Illusion of Fairness: Auditing Vulnerabilities to Distributional Manipulation Attacks

Valentin Lafargue · Adriana Laurindo Monteiro · Emmanuelle Claeys · Laurent Risser · Jean-Michel Loubes

Institut de Mathématiques de Toulouse · INRIA · IRIT · ANITI 2 · CNRS IRL CROSSING

ORAL PRESENTATION

ECML PKDD 2026

Applied Data Science Track

INSTITUT DE MATHÉMATIQUES de TOULOUSE

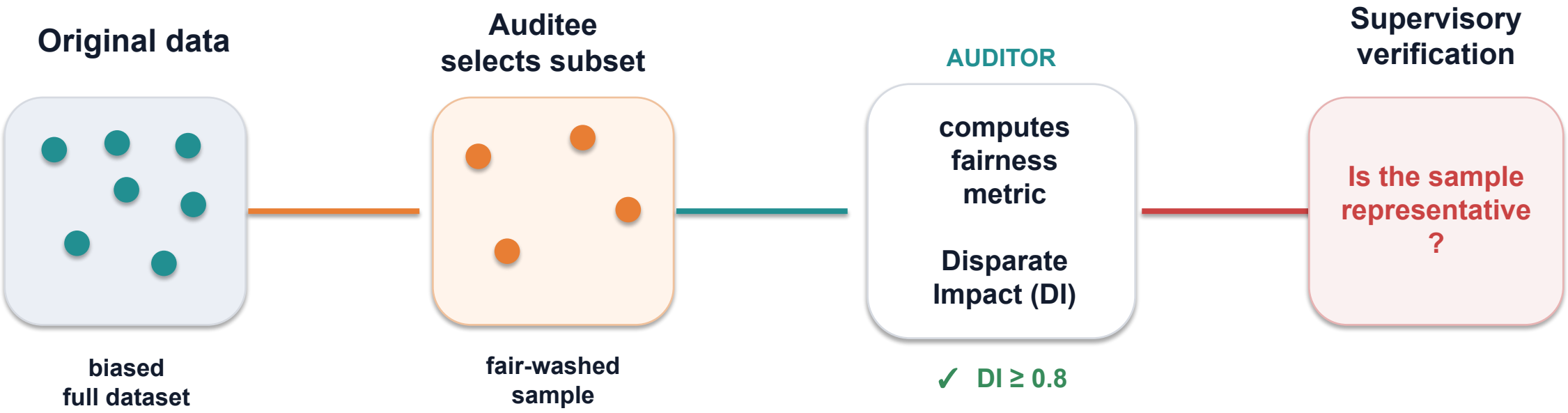
Inria

Institut de Recherche en Informatique de Toulouse
CNRS - Toulouse INP - UT - UT Capitole - UT2

ANITI Université Fédérale Toulouse Midi-Pyrénées

CROSSING
French Australian Laboratory for Humans-Automonous Agents Teaming

Can an unfair AI system pass a fairness audit?



The audit can be fooled if fairness is evaluated on a manipulated distribution.

Question of the paper

How much can an auditee improve fairness while keeping the manipulation statistically undetectable?

Three actors, three objectives

$$DI(f, \mathbb{P}, S) := \frac{\min(\mathbb{P}(\widehat{Y} = 1 \mid S = 0), \mathbb{P}(\widehat{Y} = 1 \mid S = 1))}{\max(\mathbb{P}(\widehat{Y} = 1 \mid S = 0), \mathbb{P}(\widehat{Y} = 1 \mid S = 1))}$$

Auditee

- Owns the full dataset and model, and provides the subset used by the auditor.

Goal: pass the fairness audit

Auditor

Computes the prescribed fairness metric (i.e., DI) on the disclosed subset.

Goal: assess compliance

Supervisory authority

Can access the full dataset and check whether the submitted subset is representative.

Goal: verify the auditing process

The adversarial question

Can the auditee provide sample fair enough to pass the audit, and close enough to the original data to evade supervisory detection ?

$$\operatorname{argmin}_{P \in \mathcal{P}(E), DI(f,P,S) \geq t} d(P, Q_n)$$

The paper is a cat-and-mouse game

01

Fairness can be manipulated

A biased distribution can be modified to fake compliance.

02

Manipulation can be optimized

KL and Wasserstein projections minimize distributional shift.

03

Detection is statistical

The authority tests whether the submitted sample looks genuine.

04

Some attacks still hide

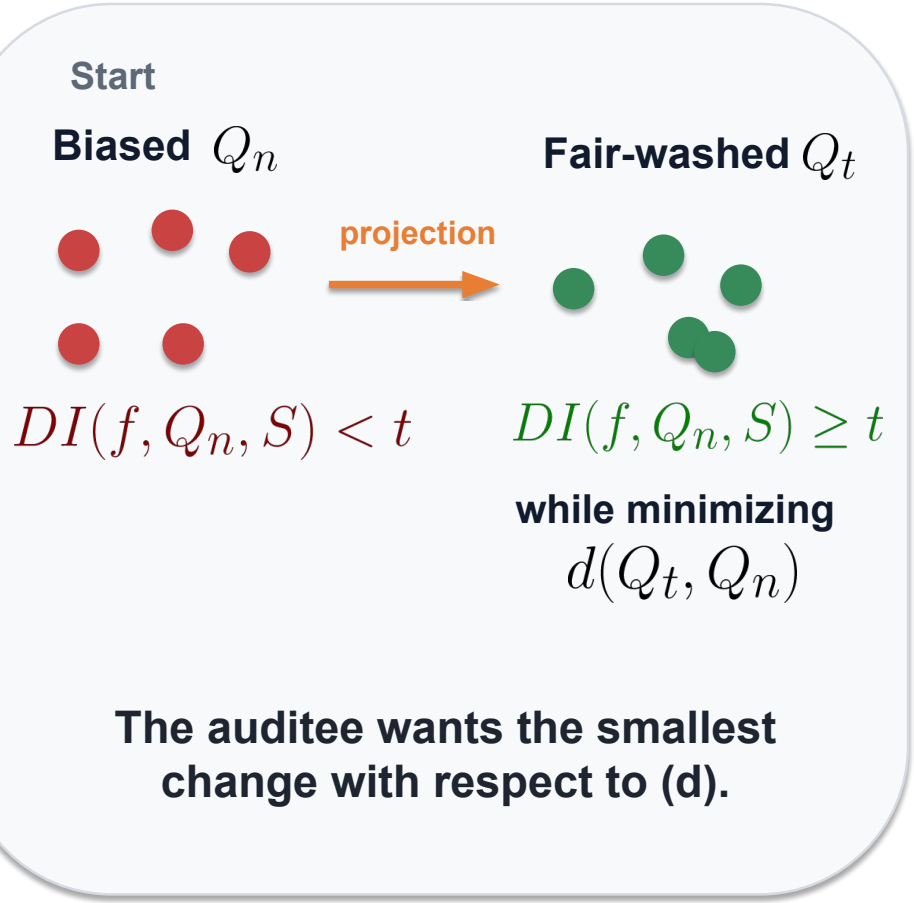
Detectability depends on test, method, dataset and sample size.

Aware of the detection strategy

The key practical question:

What should a supervisory authority actually do?

Fairwashing or preventing manipulation-proof minimum distributional movement toward fairness



Three families in the experiments

Entropic / Kullback-Leibler (extension of [1])

Reweight existing observations to meet the fairness constraint.

Continuous Wasserstein projection (new theorem)

Progressive shift toward a fairer distribution though gradient method (new observations created).

Discrete Wasserstein projection

Replacing key variables (S, $\hat{Y}$) or matching to existing individuals.

Key theoretical lens: constrained projection

Representativeness is itself a statistical hypothesis

1 · Direct hypothesis

Compare the underlying distribution $\widehat{Q}_m$ of the provided sample with the reference distribution

Kolmogorov–Smirnov test (as done in [2])

On different variables: (X,S,Ŷ) or only (S,Ŷ)

2 · Distance-based acceptance region

Generate B genuine samples of the same size m

$$d(Q_m^{(b)}, Q_n),\ b = 1, \cdots, B.$$

→ **empirical 95% threshold**

Observed submitted sample $d(\widehat{Q}_m, Q_n)$

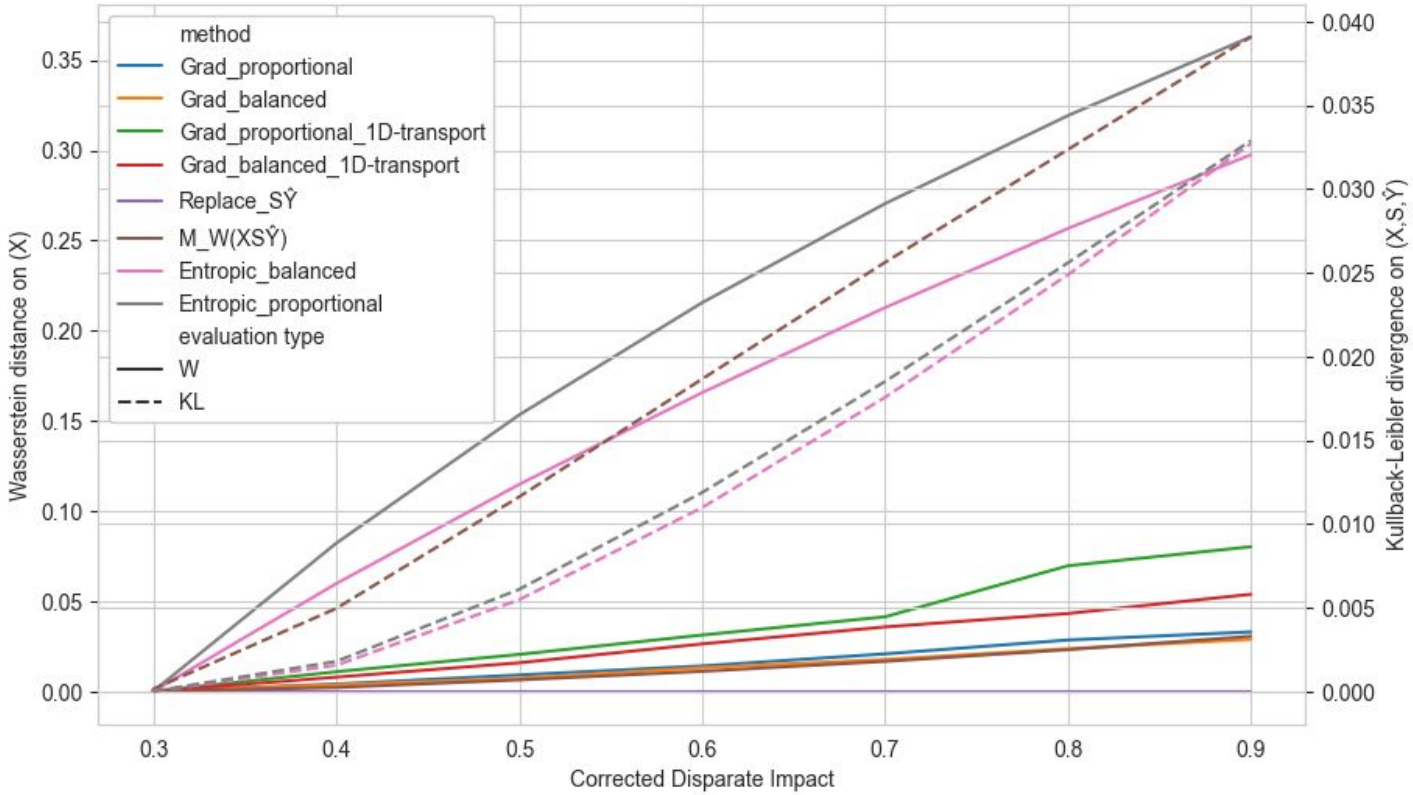
within acceptance region

implies H₀ not rejected

Crucial point

The threshold depends on the submitted sample size.

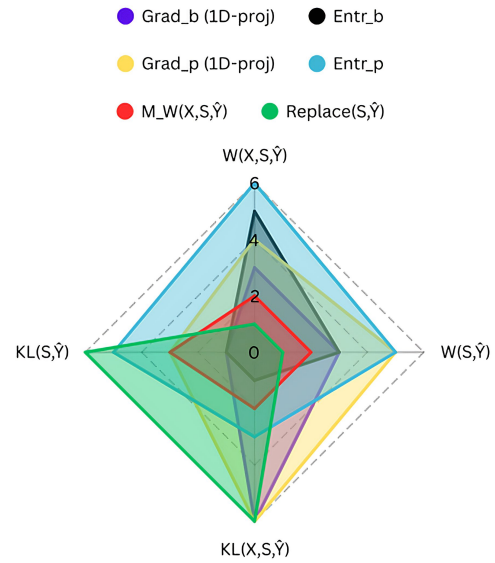
Fairwashing is a trade-off: more fairness costs more distributional shift



Fairwashing distributional cost on the Adult dataset.

What matters

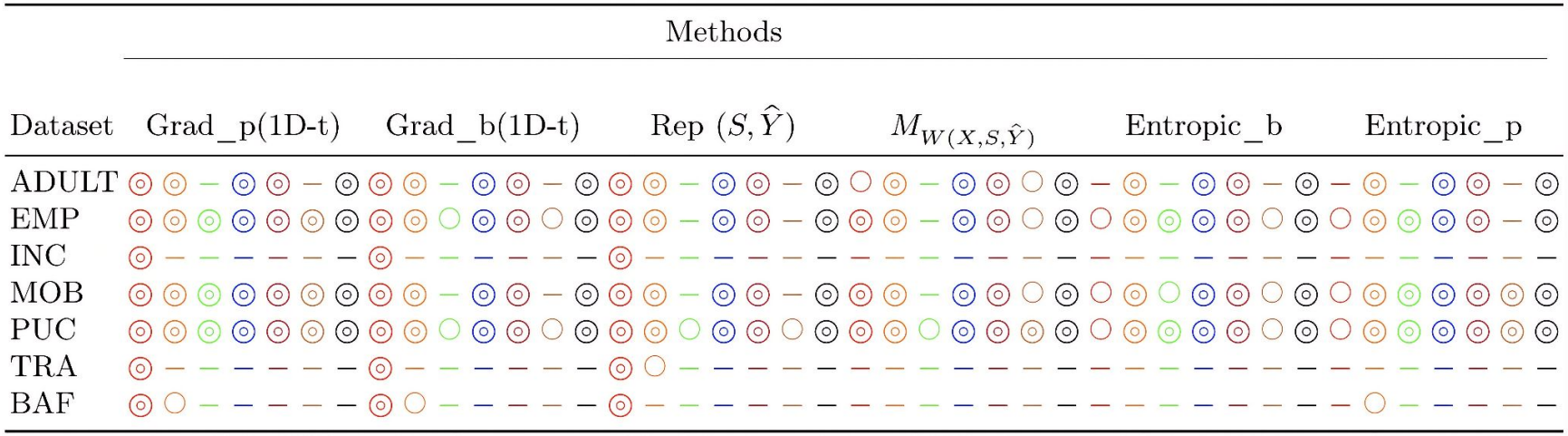
- Fairness can be increased continuously.
- Every improvement requires some distributional movement.
- The attacker therefore optimizes a fairness–detectability trade-off.



The strongest attacks exploit the test itself

Independent tests tell a consistent story

KL(X,S,Ŷ) | KL(S,Ŷ) | W(X,S,Ŷ) | W(S,Ŷ) | KS(Ŷ) | MMD(X,S,Ŷ) | MMD(S,Ŷ)



Why combine tests?

Wasserstein-optimized attacks can be caught by KL-based tests.

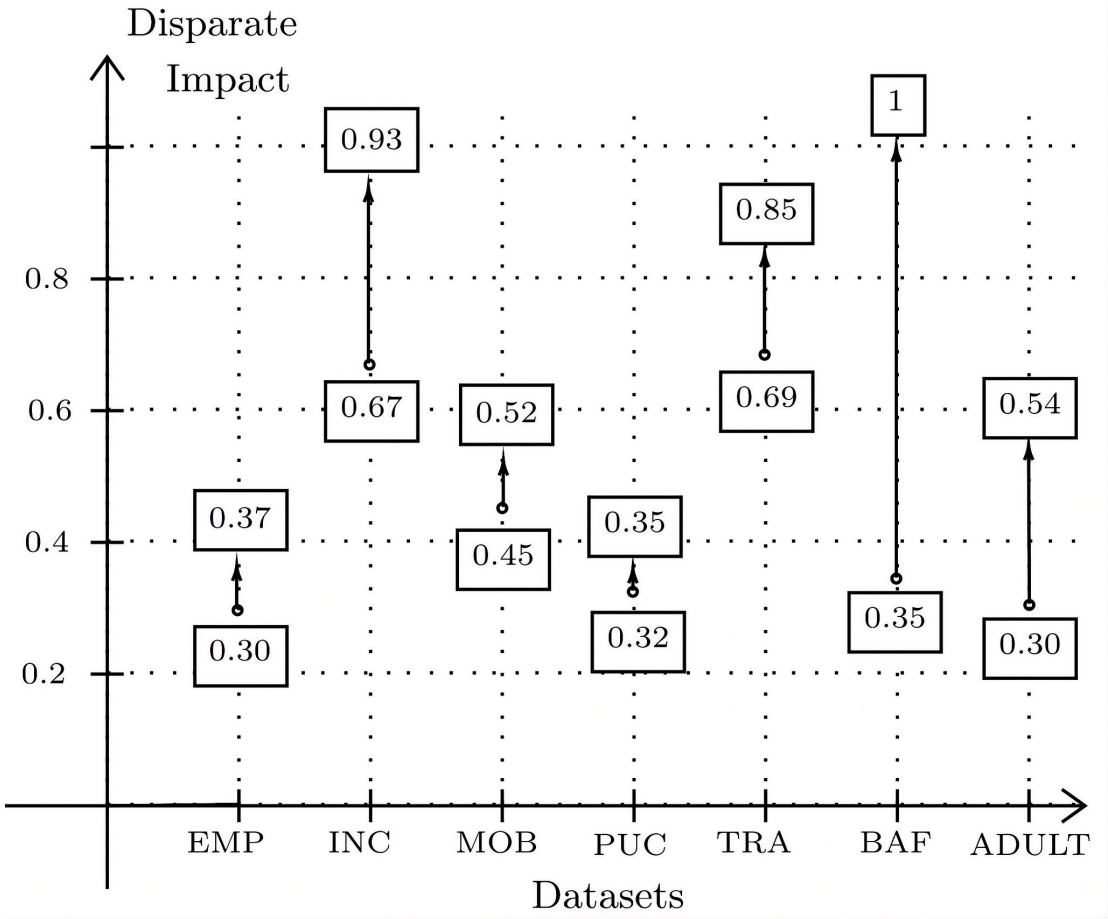
→ No single distance captures every manipulation.

Sample size is a strong lever

10% sample — more room for undetected shift

20% sample — less room for undetected shift

Detectability depends on the dataset — not just the attack



Two observations

- Combining statistical tests is an efficient detection strategy (PUC)
- The same fairwashing method is not equally detectable across datasets.
  → The original bias level does not tell the full story.

Therefore: audit design is part of the security boundary.

How should an auditor defend against fairwashing?

01

Do not audit only the fairness metric

Representativeness must be treated as a primary objective of the audit process.

02

Combine complementary statistical tests

A manipulation optimized for one distance can be exposed by another.
(Up to a certain point, adding MMD was not relevant)

03

Make the audited sample large enough

Increasing the sample size is the strongest practical lever in our experiments.

Fairness alone is not enough — the sample itself must be audited.

Job finished?

I don't think so.

What we show

Fairwashing can be optimized and sometimes remain statistically undetected.

What remains open

Designing audits that remain robust when the attacker knows the defense.

Practical constraint

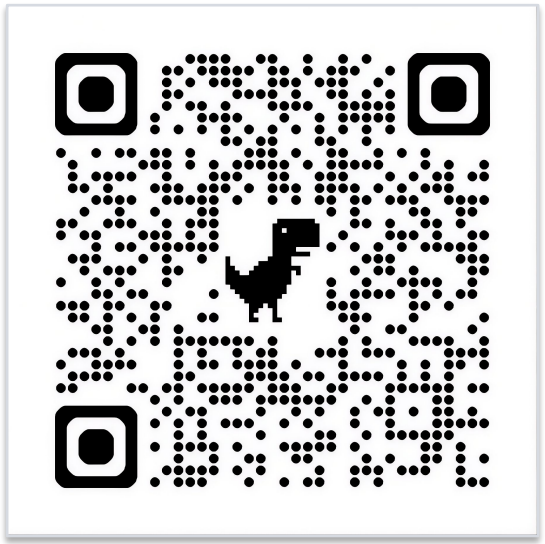
Distributional tests can become computationally expensive at scale.

The goal is not to make manipulation impossible overnight, it is to make undetectable fairwashing harder.

This is where we think the next round of audit research begins.

Thank you — and everything is reproducible

Paper · Code · Models · Datasets



GitHub

github.com/ValentinLafargue/Inspection

Hugging Face

Models:
huggingface.co/ValentinLAFARGUE/EIF-biased-classifiers

Original & manipulated datasets:
huggingface.co/datasets/ValentinLAFARGUE/EIF-Manipulated-distributions

Three final messages

- Audit the sample, not only the fairness score.
- Combine tests with complementary sensitivities.
- Prefer sufficiently large audit samples.

Questions?

[1]: Bachoc, F., Gamboa, F., Halford, M., Loubes, J.M., Risser, L.: Explaining machine learning models using entropic variable projection. Information and Inference: A Journal of the IMA, 2023.
[2]: Fukuchi, K., Hara, S., Maehara, T.: Faking fairness via stealthily biased sampling. Proceedings of the AAAI Conference on Artificial Intelligence, 2020.