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

ECML PKDD 2026
Applied Data Science Track

Can an unfair AI system pass a fairness audit?

INPUT
Full dataset Q_n

Biased; the auditee holds all of it.

AUDITEE
Selects a subset $\widehat{Q}_m$

Only this subset is disclosed.

AUDITOR
Computes Disparate Impact

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

✓ **DI ≥ 0.8** → audit passed. Sees this sample only.

Juridical / regulatory

SUPERVISORY AUTHORITY
Is the sample representative?
The only check that can catch the attack

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

Accesses the full data and the sample.

The audit can be fooled if fairness is evaluated on a manipulated distribution.
Research question: How much can an auditee improve fairness while keeping the manipulation statistically undetectable?

SAMPLE MANIPULATION: FAIRWASHING ATTACKS

Adversarial optimization problem:

$$\arg\min_{P \in \mathcal{P}(E)} d(P, Q_n) \quad \text{s.t.} \quad DI(f,P,S) \geq t$$

Reweighting — **Entropic / Kullback-Leibler projection (extension of [3])**

Reweight existing observations to meet the fairness constraint with Minimal KL divergence.

$$Q_t = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^{(t)} \delta_{X_i, S_i, \widehat{Y}_i, Y_i}$$

New points — **Continuous Wasserstein projection**

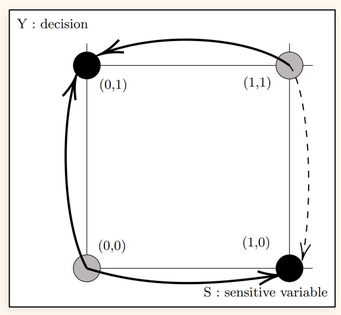
Gradient-based progressive shift of features toward a fairer distribution given a predictive model f.

Theorem. Consider $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}$. Then Q_t is a solution to $\operatorname{arginf}_{P \in \mathcal{D}_{\Phi,t}} W_2^2(P, Q_n)$ if, and only if, it is defined as $Q_t = T_{\lambda^\star} \sharp Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{T_{\lambda^\star}(Z_i)}$, where T_λ is defined as $T_\lambda(Z_i) \in \arg\min_{x \in E} \|x - Z_i\|^2 - \langle \lambda, \Phi(x) \rangle$ and $\lambda^\star$ satisfies $t = \frac{1}{n}\sum_{i=n}^{n} \Phi(T_{\lambda^\star}(Z_i))$.

Replace / Matching — **Discrete Wasserstein projection**

1. Replaces the key variables (S, Ŷ).
2. Matching method weighted by Wasserstein distance cost.

$$\frac{\mathrm{DI}(\tau_{i_0}(Z^j)) - \mathrm{DI}((Z^j))}{\|\tau_{i_0}(Z^j) - Z^j\|}$$



Let δ_0, δ_1 respectively be the increase and decrease of numerator and denominator of the DI.

To increase the DI, the changes can either be balanced $\delta_0 = \delta_1$ or proportional $\frac{\delta_0}{n_0} = \frac{\delta_1}{n_1}$.

DETECTION STRATEGY

Statistical test hypothesis: how likely is the sample from the original distribution ?

Direct hypothesis tests: K.-S. test on (X), one dimensional testing as done in [2]

Distances used: KL, Wasserstein, MMD. applied both on the full (X, S, Ŷ) and the reduced (S, Ŷ)

Distance-based acceptance region: Generate B genuine samples of size m:

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

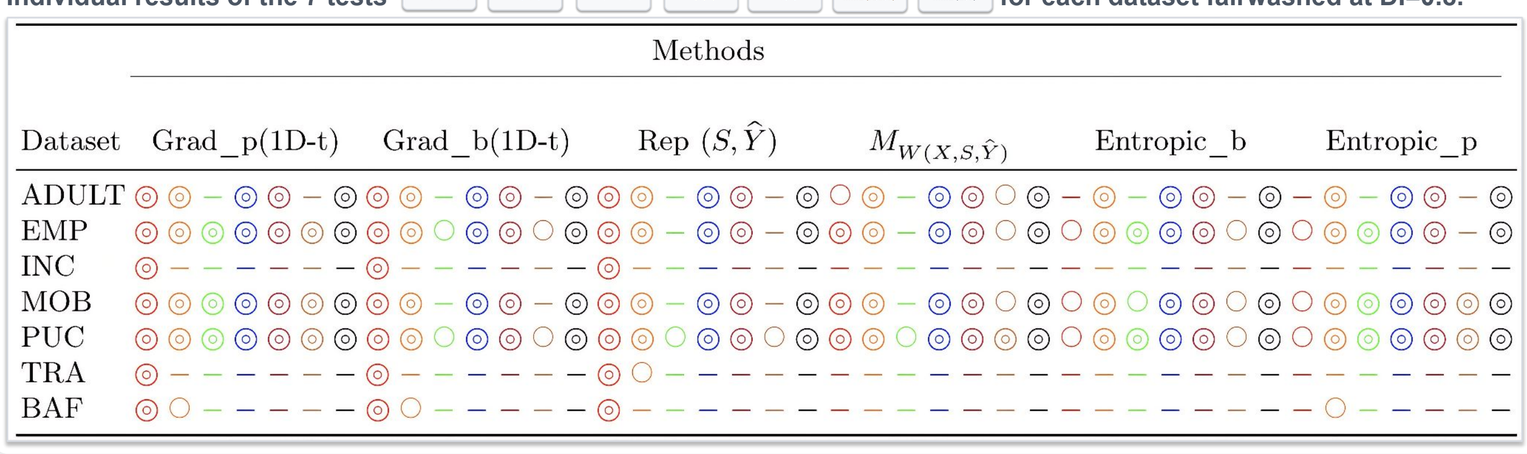
Build empirical 95% threshold for $d(\widehat{Q}_m, Q_n)$

→ Reject hypothesis if distance exceeds it.

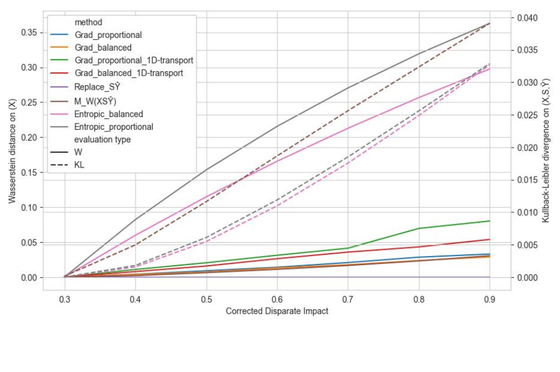
Threshold depends of sample size m: Larger m → smaller threshold → easier detection.

KEY RESULTS

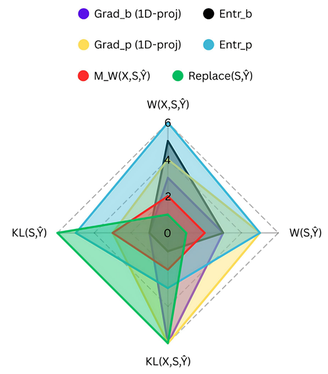
Individual results of the 7 tests KL(X,S,Ŷ) KL(S,Ŷ) W(X,S,Ŷ) W(S,Ŷ) KS(Ŷ) MMD(X,S,Ŷ) MMD(S,Ŷ) **for each dataset fairwashed at DI=0.8.**



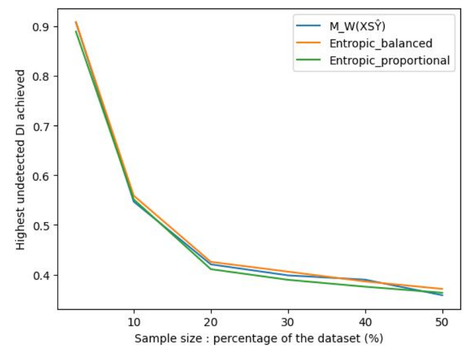
Fairness cost and detectability trade-off



Fairwashing method cost rank



Highest undetected DI per sampling size on the Adult dataset.



Highest Disparate Impact (DI) increase undetected by the combinations of 7 statistical tests, for each dataset using the different fairwashing methods. "—" means no improvement have been made using the fairwashing method on the dataset.

Dataset	Original Size (%)	Grad_p	Grad_b	Grad_p 1D	Grad_b 1D	Rep $(S, \widehat{Y})$	Entr_b	Entr_p	$M_{W(X,S,\widehat{Y})}$
ADULT	0.30	10	0.47	0.43	0.49	0.44	0.50	**0.54**	0.52	**0.54**
		20	0.39	0.40	0.38	0.39	0.41	**0.42**	0.41	**0.42**
EMP	0.30	10	–	–	–	–	–	0.36	0.36	**0.37**
		20	–	–	–	–	–	0.34	**0.36**	0.35
INC	0.67	10	0.75	–	–	–	0.88	0.94	**0.95**	0.93
		20	–	–	–	–	0.83	0.83	**0.84**	**0.84**
MOB	0.45	10	0.53	0.51	–	0.50	**0.53**	0.52	–	0.52
		20	–	–	–	0.48	0.50	0.50	–	0.50
PUC	0.32	10	–	–	–	–	–	0.33	**0.35**	**0.35**
		20	–	–	–	–	–	–	–	–
TRA	0.69	10	0.72	0.79	0.77	0.73	0.80	0.83	0.84	**0.85**
		20	–	–	–	–	0.77	0.79	0.79	**0.80**
BAF	0.35	10	–	–	–	–	–	1	1	1
		20	–	–	–	–	–	0.77	**0.80**	0.79

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

The goal is not to make manipulation impossible overnight

HOW TO MAKE FAIRWASHING HARDER?

Do not audit only the fairness metric

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

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)

Make the audited sample large enough

Increasing sample size is the strongest practical lever against undetectable shift.

CONCLUSION & OPEN QUESTIONS

- **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 strategy. This work is a step in this direction.
- **Practical constraint:** Distributional tests (especially high-dimensional Wasserstein distance) can become computationally expensive at scale.

SELECTED REFERENCES

[1] Aïvodji et al. **Fairwashing: the risk of rationalization**. ICML, 2019.
[2] Fukuchi, Hara, Maehara. **Faking fairness via stealthily biased sampling**. AAAI, 2020.
[3] Bachoc et al. **Explaining machine learning models using entropic variable projection**. Information and Inference, 2023.

Fully Reproducible

GitHub
github.com/ValentinLafargue/Inspection

Hugging Face Models
ValentinLAFARGUE/EIF-biased-classifiers

Datasets (original + manipulated)
ValentinLAFARGUE/EIF-Manipulated-distributions

Contact: valentin.lafargue@math.univ-toulouse.fr