Fairness Shields: Safeguarding against Biased Decision Makers

Filip Cano[1], Thomas A. Henzinger[2], Bettina Könighofer[1], Konstantin Kueffner[2], Kaushik Mallik[3]

[1] Graz University of Technology, [2] Institute of Science and Technology Austria, [3] IMDEA Software Institute

AAAI-25 / IAAI-25 / EAAI-25

AI-based decision-makers are prone to reproducing existing social bias in their decisions, with respect to features such as race or gender. This motivates the existence of methods to enforce *fairness*.

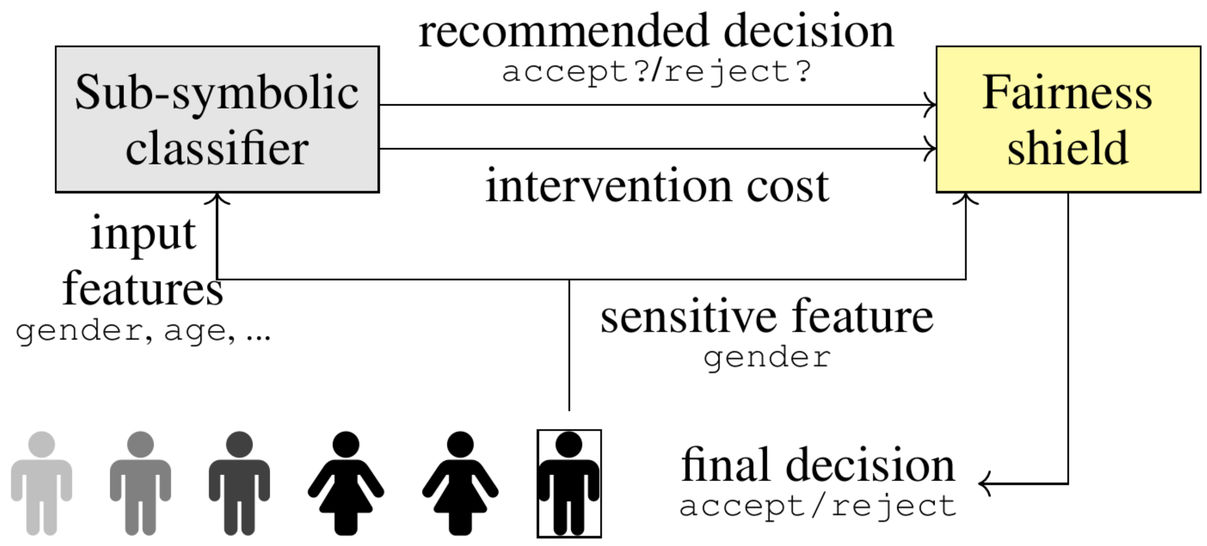
- **Long Term vs Short Term Fairness.** Most methods in the literature [1] guarantee fairness in the long run. However, a decision-maker can be fair on average and still produce biased runs in specific intervals.
- **Fairness Shielding.** We introduce *fairness shields* to guarantee fairness in finite runs of a given length. We also study under which conditions short-term guarantees lead to long-term guarantees.

Why shielding?

- Shields are designed to monitor execution and act only when required to satisfy the fairness constraint.
- Shields are agnostic to the shielded classifier. A shield can be applied to *complement* an already fair classifier, so it ensures fairness in the few cases the original classifier would not. It can also be applied to a black-box classifier.

[1] Barocas *et al.* "*Fairness and Machine Learning*". MIT Press (2023).

Overview



Formal Model

- **ML classifier**. Formally, the ML classifier and the population distribution form a distribution of the shield's input, denoted by $\theta\colon \mathcal{G} \times \mathcal{F} \to \mathcal{X}$. For an input (g, f), where g is the *group membership* (or sensitive feature), and f is the rest of the features, $\theta(g, f) = x = (g, r, c)$, where r is the *recommendation* given by the ML classifier, and c is the *cost* of changing the recommendation.
- **Shield**. Formally, the shield is a function $\pi\colon (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \mathcal{Y}$. For an input $(x_1, y_1), \ldots, (x_n, y_n), x_{n+1}$, where x_i is the ML classifier's output for the i-th instance, and $y_i \in \{0, 1\}$ is the final decision of the shield for the i-th instance; the shield produces a decision y_{n+1}.
- **Feasible traces**. A trace $\tau \in (\mathcal{X} \times \mathcal{Y})^T$ is *feasible* with shield π if there exists an input sequence $x_1, \ldots, x_T \in \mathcal{X}^T$, such that when applying the shield's decisions, yields τ. We call this $\mathtt{FT}^T_{\theta,\pi}$.

Fairness metrics

In this work, we focus on *group fairness* (in contrast to other notions of fairness, individual fairness). Given a trace $\tau \in (\mathcal{X} \times \mathcal{Y})^*$:

- **Welfare function**. A welfare function $\mathtt{WF}^g(\tau)$ measures how "good" group g is doing in trace τ. For example, the ratio of accepted individuals is a welfare function.
- **Fairness metric**. A fairness metric is the difference of welfare functions among different groups: $\varphi(\tau) = |\mathtt{WF}^a(\tau) - \mathtt{WF}^b(\tau)|$. For ratio of accepted individuals, the metric is *demographic parity*.
- **Fair traces**. Given a fairness metric φ and a *threshold* κ, a trace is *fair* if $\varphi(\tau) \leq \kappa$.
- **Fairness enforcement**.
  - **Finite Horizon**. A shield π enforcer fairness with *finite horizon* T if for all trace $\tau \in \mathtt{FT}^T_{\theta,\pi}$, $\varphi(\tau) \leq \kappa$.
  - **Periodic shielding**. A shield π enforcer fairness with *periodic horizon* T if for all $k \geq 0$ and for all trace $\tau \in \mathtt{FT}^{kT}_{\theta,\pi}$, $\varphi(\tau) \leq \kappa$.

Shield Synthesis

The *optimal shield* minimizes expected cost among shields that only produce fair traces up to length T, denoted $\Pi^{\theta,T}_{\mathtt{fair}}$.

$$\pi^* = \arg\min_{\pi \in \Pi^{\theta,T}_{\mathtt{fair}}} \mathbb{E}[\mathit{cost}; \theta, \pi, T].$$

To synthesize π^*, we compute $v(\tau)$ recursively:

$$v(\tau) = \min_{\pi \in \Pi^{\theta,(T-|\tau|)|\tau}_{\mathtt{fair}}} \mathbb{E}[\mathit{cost} \mid \tau; \theta, \pi, T - |\tau|].$$

- **Base case.** For $|\tau| = T$: $\quad v(\tau) = \begin{cases} 0 & \varphi(\tau) \leq \kappa, \\ \infty & \text{otherwise.} \end{cases}$
- **Recursive case.**

$$v(\tau) = \sum_{x=(g,r,c) \in \mathcal{X}} \theta(x) \cdot \min\left\{v(\tau, (x, y = r)), v(\tau, (x, y \neq r)) + c\right\}.$$

Types of Periodic Shields

Our synthesis method only guarantees fair traces of length T. To produce periodically fair traces, we have three alternatives:

- **Static-fair.** Re-use π^* for longer traces.
  - Works well experimentally. The formal guarantees on fairness almost never apply.
- **Static-BW.** Modify the shield synthesis method to enforce $\mathtt{WF}^g(\tau) \in [l, u]$, for certain bounds l, u.
  - Reusing Static-BW shields has stronger fairness guarantees. These guarantees are only lost for traces that are significantly skewed in favour of one group or the other.
  - In practice, they produce overly conservative enforcement, incurring high intervention costs and utility loss.
- **Dynamic.** Synthesize a new shield after every T instances, modifying the fairness property φ to take into account the already-seen trace. Best performance, but most expensive.

Experimental Evaluation

Experimental Setup

- *Datasets*: Adult, COMPAS, German Credit, Bank Marketing.
- *ML algorithms*: DiffDP, HSIC, LAFTR, PRemover, ERM.
- *Fairness metric*: Dem. Parity ($T = 100$), Equal Opportunity ($T = 75$).

	FinHzn					Periodic		
adult, gender	0.43	1.90	0.53	1.56	0.44	3.44	11.85	1.36
	2.45	1.19	1.96	1.61	1.37	4.83	7.98	0.70
bank, age	7.43	8.73	6.88	6.50	7.70	6.86	6.01	1.61
	1.10	1.77	1.23	-0.23	1.48	3.93	7.53	0.69
compas, race	1.45	0.96	0.60	0.92	0.99	1.99	8.35	0.51
	1.57	3.51	0.63	2.87	3.19	8.53	8.95	2.11
german, gender	0.28	0.49	1.00	0.95	1.37	3.73	9.55	0.76
adult, gender	8.54	11.73	8.62	11.27	8.20	11.30	6.45	7.02
	9.34	11.57	11.68	10.84	10.18	12.52	6.50	10.10
bank, age	1.64	2.96	3.34	1.72	2.35	2.95	3.16	2.34
	16.41	18.40	19.20	19.79	16.51	21.59	9.51	7.97
compas, race	17.84	18.93	20.27	17.99	17.56	20.77	10.66	5.48
	59.05	58.68	59.11	59.56	60.46	62.08	9.13	14.29
german, gender	53.46	54.44	52.44	52.28	53.69	61.57	10.34	9.08
	DiffDP	ERM	HSIC	LAFTR	PRemover	Static-Fair	Static-BW	Dynamic
	ML Algorithm					Shield		

Table 1. Utility loss (in %) incurred by FinHzn shields for different ML models (left) and by periodic shields on the ERM model (right) for the fairness properties DP (top, green) and EqOpp (bottom, blue). Lighter colors indicate smaller utility loss.

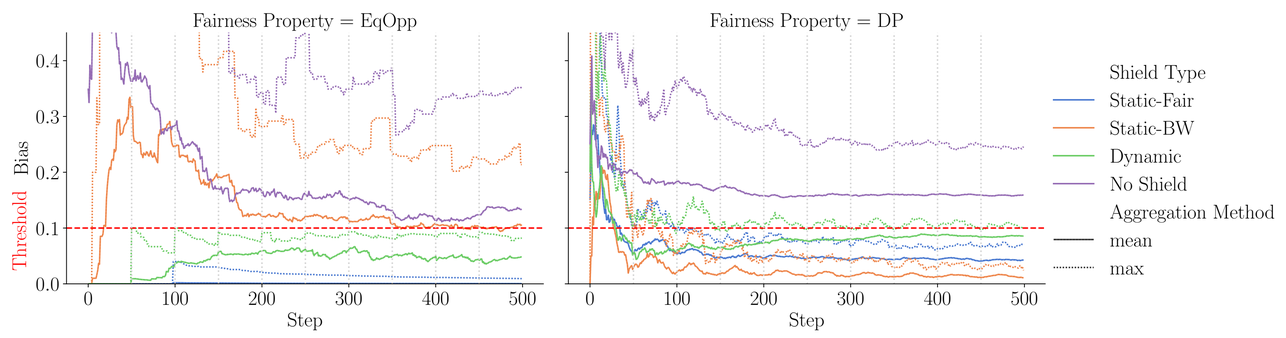


Figure 1. Variations of bias over time for the ERM classifier on the Adult dataset.

		Recomputation	Assumption satisfied	Fairness satisfied
DP	Static-Fair	no	0.0%	95.71%
	Static-BW	no	43.8%	83.1%
	Dynamic	yes	100%	100%
EqOpp	Static-Fair	no	0.0%	100%
	Static-BW	no	4.1%	56.4%
	Dynamic	yes	49.8%	100%

Table 2. Comparison of different types of fairness shields.

Scan to read the full paper