Balancing the Scales

Reinforcement Learning for Fair Classification

Leon Eshuijs, Shihan Wang, Antske Fokkens

LREC 2026 Palma

Identity-Aware AI 2026 - LREC workshop

TL;DR

1. We adapt three RL algorithms (LinUCB, DQN, PPO) for fair classification using a Contextual Multi-Armed Bandit framework with reward scaling.
2. LinUCB excels on binary tasks with under 2 epochs of training, while DQN and PPO are stronger on multi-class problems.
3. Algorithmic fairness methods cut the GAP roughly in half compared to representational debiasing, while keeping accuracy competitive.

Why *fairness* needs more than representation debiasing

Most fairness work removes bias from representations. But representational fairness ≠ empirical fairness (Shen et al. 2022). We shape the training signal directly, optimizing fairness without throwing away useful information.

Contextual Multi-Armed Bandit (CMAB)

We cast fair classification as a CMAB framework:

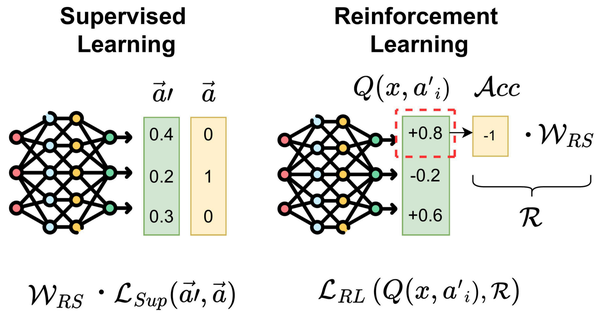
- Input context vector: x_t
- Chose an action: a_t
- Receives a reward: r_t
- Protected attribute: g_t

Acc ∈ {+1, -1}
W : reward scales

Reward function

$$R(a, a_{pred}, g) = \mathrm{Acc}(a, a_{pred}) \cdot \mathcal{W}(a, g)$$

Three RL algorithms + Supervised Learning



LinUCB Classical CMAB - Linear Policy

Linear reward model with confidence bounds for exploration. One arm per class with its own weight vector

DQN-bandit Original MDP - Deep RL - Value-based

Learns a Q-function for state-action pairs. Adapted to a one-step MDP with ε-greedy exploration.

PPO-bandit Original MDP - Deep RL - Policy gradient

Directly optimises the policy with clipped updates and an entropy bonus to encourage exploration.

Four types of Scaling

Reduce majority

Extends Lin et al. 2020

$$\mathcal{W}_{\rho+}(a, g) = \begin{cases} 1 & \text{minority} \\ \rho^a_{imb} & \text{majority} \end{cases}$$

Boost minority

Ours

$$\mathcal{W}_{\rho-}(a, g) = \begin{cases} (\rho^a_{imb})^{-1} & \text{minority} \\ 1 & \text{majority} \end{cases}$$

Equal Opportunity

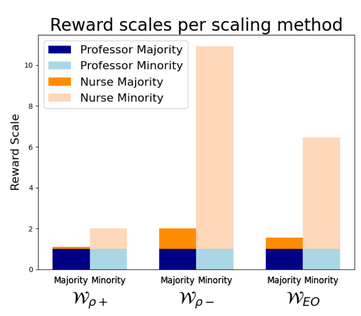
Best theoretical foundation

$$\mathcal{W}_{EO}(a, g) = \frac{1}{2} \frac{1}{P(g|a)}$$

Inverse Prob. Weighting

Höfler et al. 2005

$$\mathcal{W}_{IPW}(a, g) = \frac{1}{P(a, g)}$$

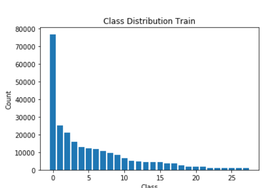


Two types of imbalance

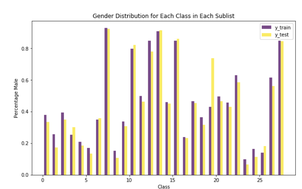
1. Class imbalance based on the prediction class
2. Sub-Class imbalance of the protected attribute

→*What we focus on*

Class distribution



Sub-Class distribution



Experimental Setup

Datasets:

- **BiasBios** (*De-Arteaga et al., 2019*):
  - 393k biographies, 28 professions
  - gender as protected attribute
- **Emoji** (*Elazar and Goldberg, 2018*):
  - Binary sentiment classification on Twitter
  - race (AAE/SAE) as protected attribute

Fairness Metrics:

The Root Mean Square of the difference in True Positive Ratio (TPR) between protected groups.

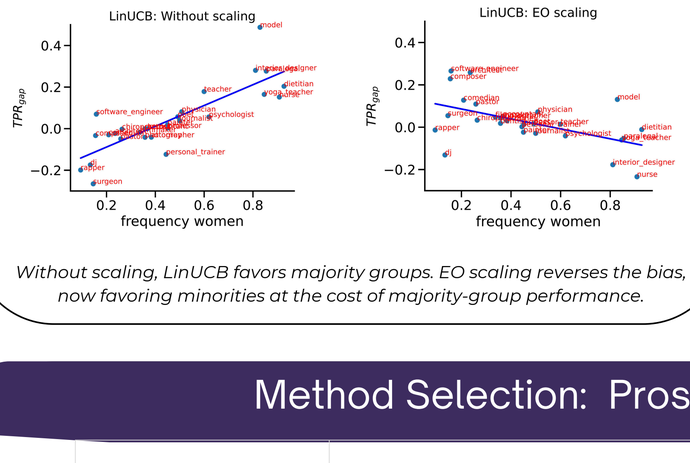
$$\mathrm{GAP} = \sqrt{\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left(\mathrm{TPR}^a_{gap}\right)^2}$$

Main Results:

Algorithm		BiasBios — 28 Classes Acc ↑	GAP ↓	DTO ↓	F1 ↑	Time ↓	Emoji — 2 Classes Acc ↑	GAP ↓	DTO ↓	Time ↓
Sup	VANILLA	81.0 ± 0.1	16.4 ± 0.5	9.3	73.8 ± 0.3	1.0	72.3 ± 0.1	38.1 ± 0.6	28.3	1.0
INLP	REPR	80.2 ± 0.6	9.7 ± 0.4	2.8	71.7 ± 1.4	50.1	63.5 ± 3.6	24.1 ± 5.4	18.6	3.6
MP	REPR	**81.1 ± 0.1**	13.9 ± 0.6	6.8	**74.0 ± 0.2**	2.6	71.8 ± 0.3	17.1 ± 1.0	8.1	2.3
Sup^{EO}	OURS	80.1 ± 0.2	**7.1 ± 0.5**	**1.0**	71.7 ± 0.5	1.0	**75.5 ± 0.1**	11.4 ± 1.1	1.4	1.0
LinUCB^{EO}	OURS	74.6 ± 0.2	12.2 ± 0.5	8.3	59.8 ± 1.1	31.9	75.3 ± 0.2	10.4 ± 0.7	**0.5**	2.8
$\mathrm{DQN}^{EO}_{bandit}$	OURS	79.2 ± 0.1	10.1 ± 0.4	3.6	66.4 ± 0.2	57.4	70.8 ± 0.8	**10.0 ± 1.0**	4.8	30.2
$\mathrm{PPO}^{EO}_{bandit}$	OURS	79.2 ± 0.2	8.5 ± 0.2	2.4	66.0 ± 0.8	2.9	75.4 ± 0.1	14.4 ± 0.6	4.4	3.0

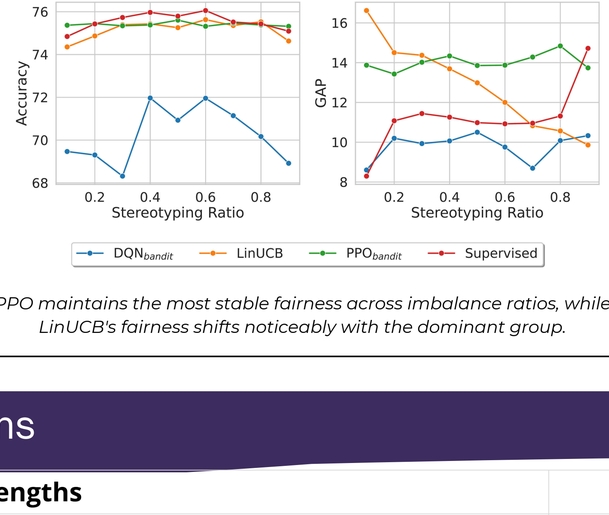
- Supervised + EO leads on fairness; PPO is the strongest RL method.
- LinUCB hits the best fairness-accuracy trade-off in fewer than 2 epochs.
- Algorithmic methods obtain much lower GAP than representational methods

Reward scaling can overshoot



Without scaling, LinUCB favors majority groups. EO scaling reverses the bias, now favoring minorities at the cost of majority-group performance.

Sensitivity to Imbalance



PPO maintains the most stable fairness across imbalance ratios, while LinUCB's fairness shifts noticeably with the dominant group.

Effect of Representational Bias

Algo+$\mathcal{W}^{EO}$	Explicit Gender Info Accuracy ↑	GAP ↓	MP-Debiased Accuracy ↑	GAP ↓
Sup	**80.2** (+ 0.1)	**7.2** (+ 0.1)	**80.0** (- 0.1)	**7.4** (+ 0.3)
LinUCB	74.5 (+ 0.1)	11.7 (- 0.5)	74.3 (- 0.3)	11.5 (- 0.7)
DQN_{bandit}	79.2 (+ 0.0)	10.0 (- 0.1)	79.0 (- 0.2)	8.6 (- 1.5)
PPO_{bandit}	79.3 (+ 0.1)	8.7 (+ 0.2)	79.2 (+ 0.0)	9.7 (+ 1.2)

Manipulating the gender signal in the embeddings has minimal effect on fairness

Method Selection: Pros and Cons

	Strengths	Limitations
LinUCB	Trains in <2 epochs · only 1 hyperparameter (α) · best on binary	Scales poorly with class count · over-corrects on multi-class with scaling
DQN-bandit	Strong without scaling · handles sub-class imbalance naturally with enough data	Slowest training (replay buffer) · misses very sparse classes
PPO-bandit	Most stable across stereotyping ratios · training time close to supervised	Misses very sparse classes · sensitive to hyperparameters (entropy, clip)
Sup + EO	Best multi-class fairness · simplest to implement · fastest	Doesn't leverage RL benefits (e.g., non-differentiable objectives)

Code:

https://github.com/watermeleon/RL_for_imbalanced_classification

Contact:

l.eshuijs@vu.nl

Funding:

Hybrid Intelligence Centre · NWO

VU Amsterdam · Utrecht University