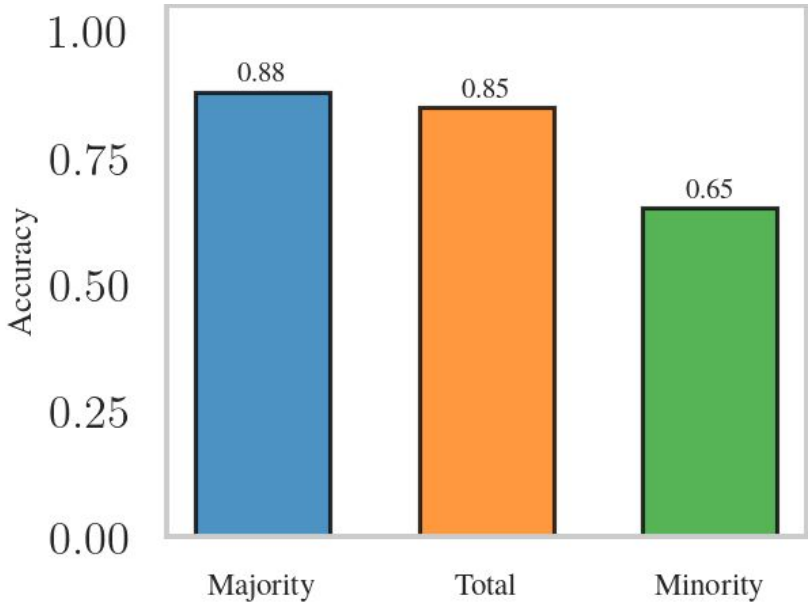
Fairness for the People, by the People: Minority Collective Action

Omri Ben-Dov[1], Samira Samadi[1], Amartya Sanyal[2], Alexandru Ţifrea[3]

[1]Max Planck Institute for Intelligent Systems, Tübingen AI Center [2]Department of Computer Science, University of Copenhagen [3]ETH Zurich

Fairness in machine learning

Features x_i , labels y_i

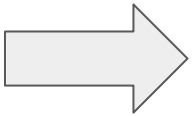


Group fairness
similar outcomes to different groups

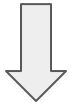
Solutions for fairness

Change the learning algorithm:

- Edit the data
- Modify the training objective
- Post-process the results
- …



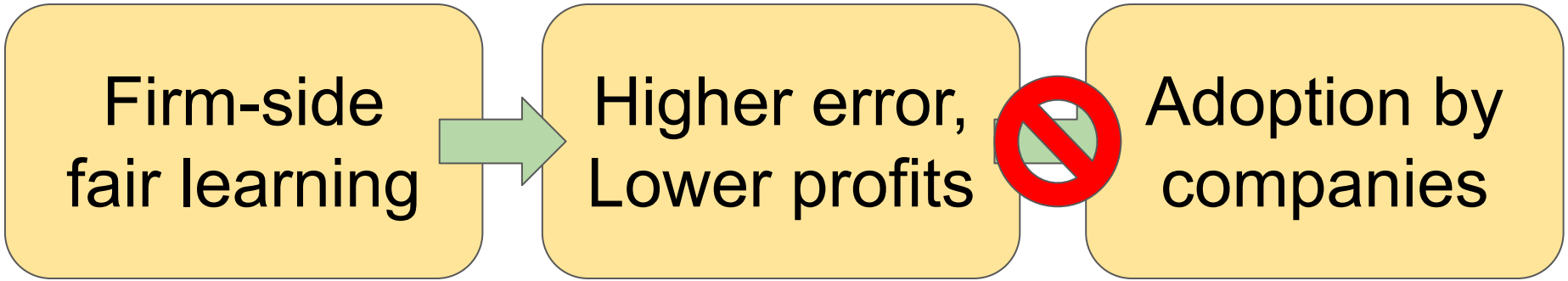
Fairness improves



Lower predictive accuracy

Menon, A.K. and Williamson, R.C., 2018, January. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency* (pp. 107-118). PMLR.

Solutions for fairness



When a company learns from user-data,
can a minority induce fairness?

Algorithmic Collective Action

Hardt, M., Mazumdar, E., Mendler-Dünner, C. and Zrnic, T., 2023, July. Algorithmic collective action in machine learning. In *International Conference on Machine Learning* (pp. 12570-12586). PMLR.

Assume a classifier trained on data from end-users:

- One person will have a negligible effect on the classifier.
- By collaborating, people can influence the classifier.
- How many and to what goal?

Algorithmic Collective Action

Hardt, M., Mazumdar, E., Mendler-Dünner, C. and Zrnic, T., 2023, July. Algorithmic collective action in machine learning. In *International Conference on Machine Learning* (pp. 12570-12586). PMLR.

Goal: Make the classifier ignores a signal *g*

Success: $$S(\alpha) = \mathbb{P}_0\left[h(g(x)) = h(x)\right]$$

Collective size — Signal to "ignore" — Classifier

Action:
Change labels

$$x, y \to x, \underset{y' \in \{0,1\}}{\operatorname{argmax}}\, \mathbb{P}_0\left(y' | g(x)\right)$$

Algorithmic Collective Action

Hardt, M., Mazumdar, E., Mendler-Dünner, C. and Zrnic, T., 2023, July. Algorithmic collective action in machine learning. In *International Conference on Machine Learning* (pp. 12570-12586). PMLR.

Goal: Make the classifier ignores a signal *g*

Success: $$S(\alpha) = \mathbb{P}_0\left[h(g(x)) = h(x)\right]$$

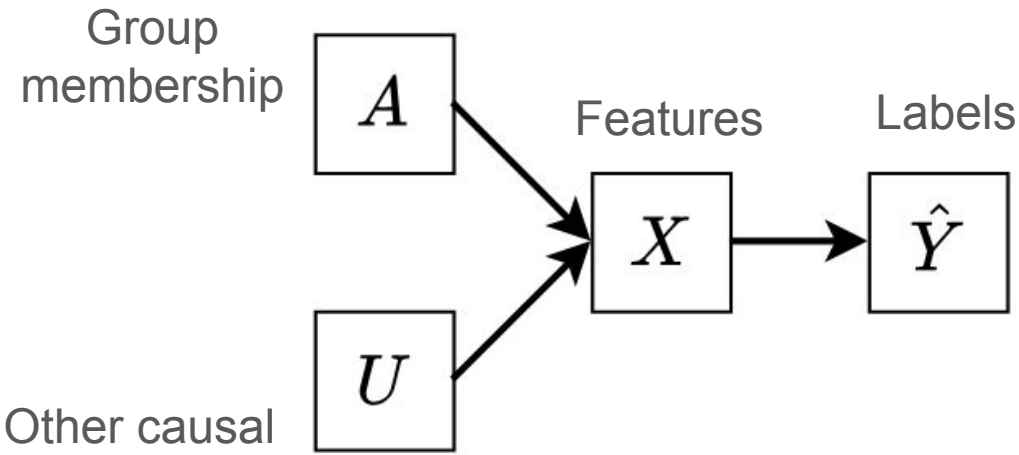
Collective size — Signal to "ignore" — Classifier

Action: Change labels $$x, y \to x, \operatorname{argmax}_{y' \in \{0,1\}} \mathbb{P}_0(y' | g(x))$$

If *g* contains group membership, the classifier will "ignore" group bias.

Counterfactual Fairness

Assume a causal model



Counterfactual signal

$$g(x) = x_{A \leftarrow \text{Majority}}$$

Minority Collective Action

Cannot accurately modify all labels:

- We do not know the counterfactuals
- Only the minority may participate

$$x, y \to x, \underset{y' \in \{0,1\}}{\operatorname{argmax}} \mathbb{P}_0\left(y' | g\left(x\right)\right)$$

Method:

1. Find “negative” minority members
2. Estimate the likelihood of a counterfactual “positive” label

Minority Collective Action

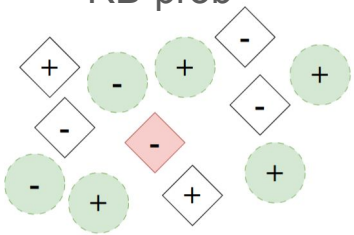
Cannot accurately modify all labels:

- We do not know the counterfactuals
- Only the minority may participate

Method:

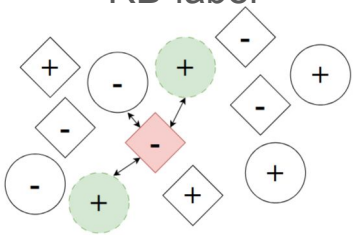
1. Find “negative” minority members
2. Estimate the likelihood of a counterfactual “positive” label

RB-prob



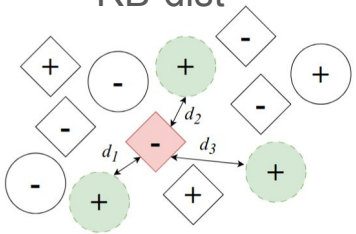
$$s_i = f(x_i)$$

RB-label



$$s_i = \sum_{j \in K_i} \mathbf{1}\{y_j = 1\}$$

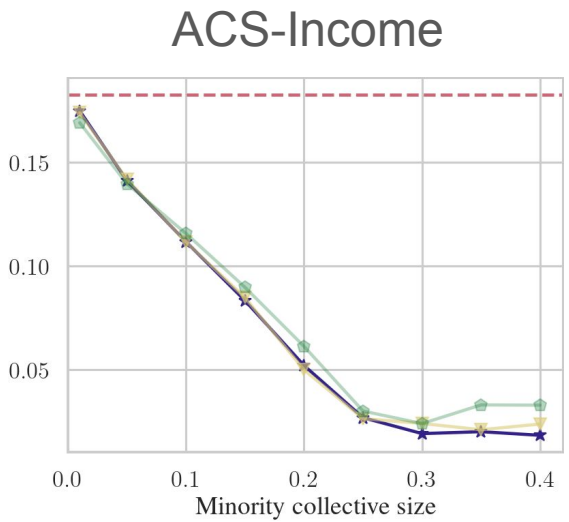
RB-dist

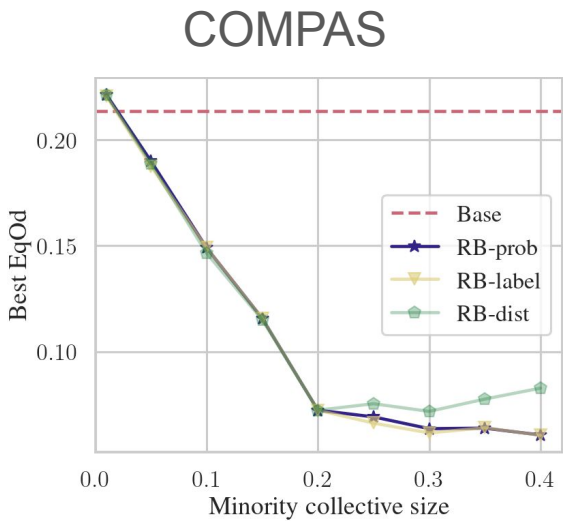


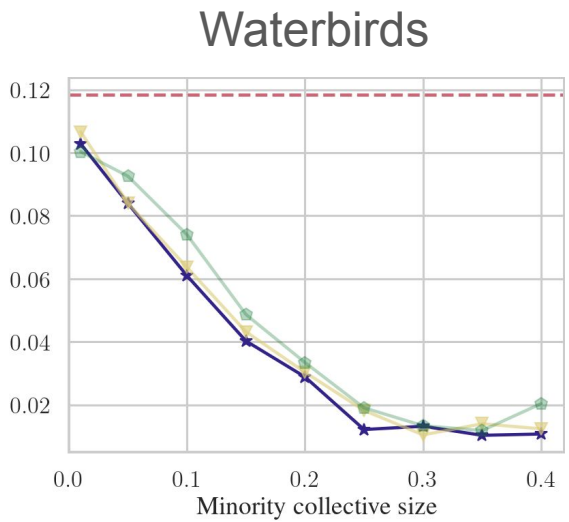
$$s_i = -\frac{1}{k} \sum_{j \in K_i} \|x_i - x_j\|_2$$

3. Top candidates flip their labels

Minority Collective Size

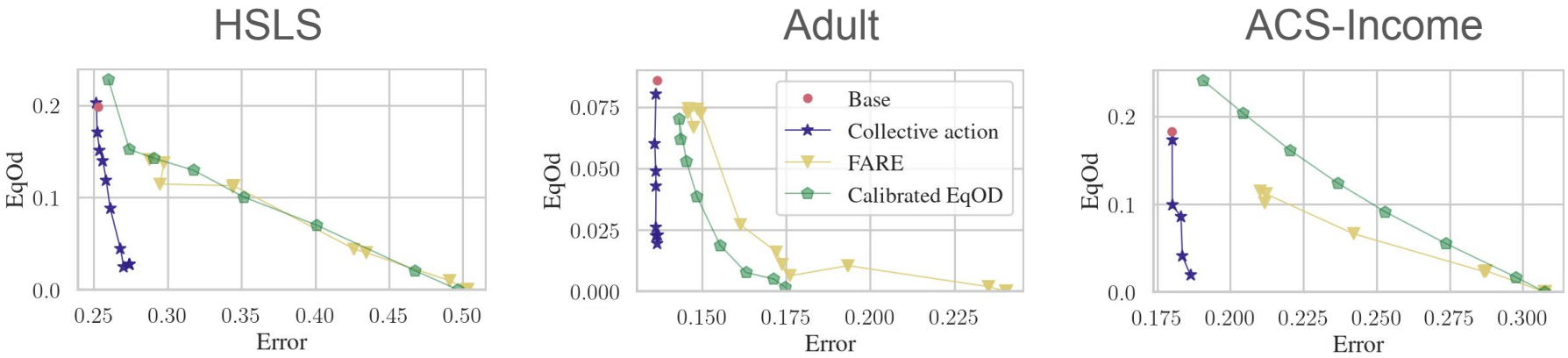






20–30% of the minority attains the least fairness violation.

Compare With Firm-Side Methods



Unlike firm-side FARE and calibrated equalized odds,
a minority cannot get perfect fairness, but adds smaller error.

Jovanović, N., Balunovic, M., Dimitrov, D.I. and Vechev, M., 2023, July. Fare: Provably fair representation learning with practical certificates. In *International Conference on Machine Learning* (pp. 15401-15420). PMLR.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. and Weinberger, K.Q., 2017. On fairness and calibration. *Advances in neural information processing systems*, *30*.

Conclusion

1. Collective action framework for fairness.
2. Fairness improves linearly with collective size until around 30%.
3. Generally, a minority cannot achieve perfect fairness.
4. Smaller error than firm-side fairness methods.

More details and experiments in the preprint https://arxiv.org/abs/2508.15374