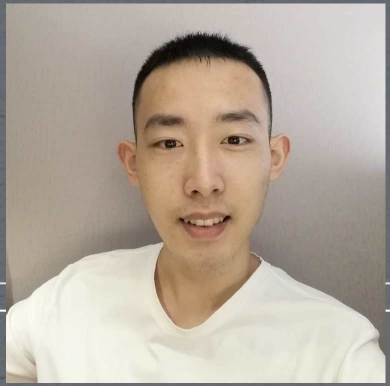
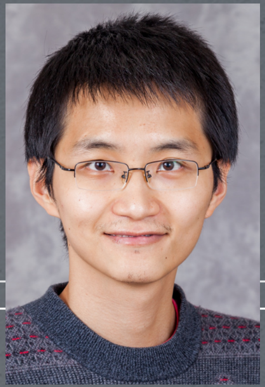
Fair for All: Best-effort guarantees for Fairness in Classification

Anilesh Krishnaswamy (Postdoc @Duke CS), joint work with:



Zhihao Jiang
(Tsinghua)

Kangning Wang
(Duke)

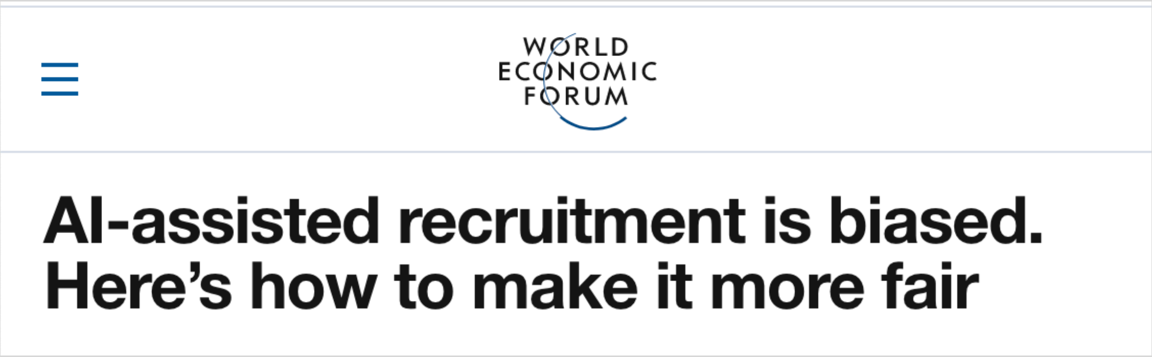


Yu Cheng
(UIC)

Kamesh Munagala
(Duke)

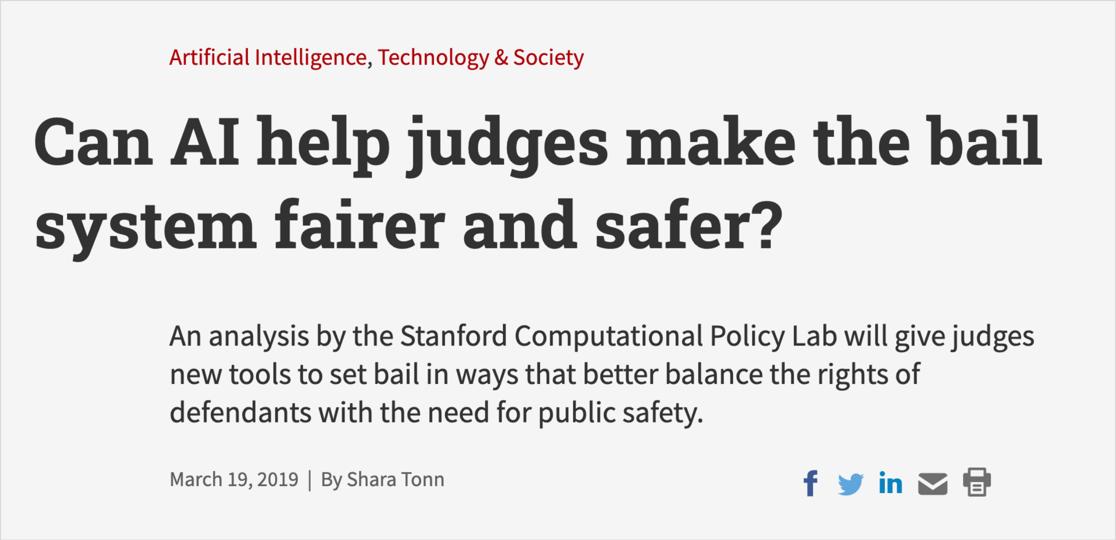
AI/algorithms are making critical decisions about individuals

- Will you get **a loan**?
- Will you get **a job**?
- Will you get **out on bail**?



WORLD ECONOMIC FORUM

AI-assisted recruitment is biased. Here's how to make it more fair

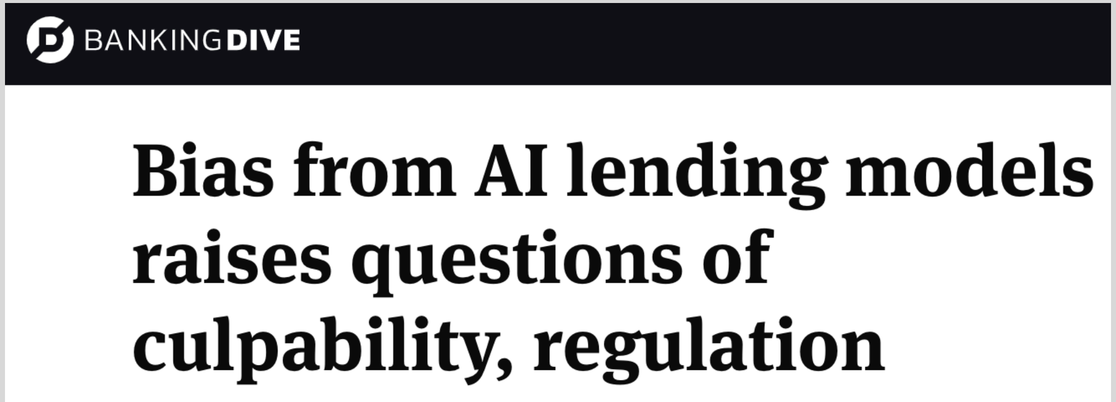


Artificial Intelligence, Technology & Society

Can AI help judges make the bail system fairer and safer?

An analysis by the Stanford Computational Policy Lab will give judges new tools to set bail in ways that better balance the rights of defendants with the need for public safety.

March 19, 2019 | By Shara Tonn



BANKING DIVE

Bias from AI lending models raises questions of culpability, regulation

Standard notions of fairness

- Many well-motivated statistical notions of fairness: *parity, equalized odds*, etc.
- Common theme: For a given list of protected groups, seek to equalize some statistical measure across them.
  - e.g. *parity* – equalize accuracy across groups.
- However, there are some drawbacks to such an approach.
  - **Dependence on the specification of groups**: mis-specifying or mis-calibrating them could potentially harm some sections.
  - **Aiming for an absolute guarantee**: some of the groups could be inherently harder to classify than others, trying to achieve parity could do more harm than good by bringing down the accuracy on a group that is easier to classify.

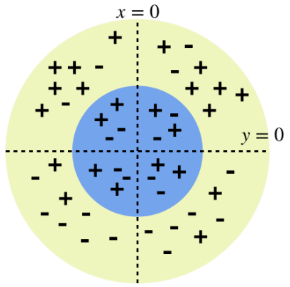


Figure 1: Given two groups Blue and Yellow (with true labels as shown), we have to choose just between the two classifiers $x = 0$ and $y = 0$. The Blue group is inherently harder to classify. Equalized odds makes us choose the classifier $x = 0$, thereby hurting the Yellow group. We could choose $y = 0$ with no aggregate effect on Blue, doing much better on Yellow.

Best-effort guarantees

- We aim for *"best-effort"* fairness guarantees:
  - Relative to how well each group can be classified in itself.
- Moreover, we want these guarantees to hold for broad classes of groups.
  - In particular, we apply our fairness notion to:
    a) All possible groups in the data
    b) More streamlined classes of groups, such as all linearly separable ones.
- Deterministic classifiers cannot be robust to large classes of groups – we focus on randomized classifiers (randomized over a given hypothesis space).
- We analyze both the above-mentioned settings and provide efficient algorithms for each.

All possible groups

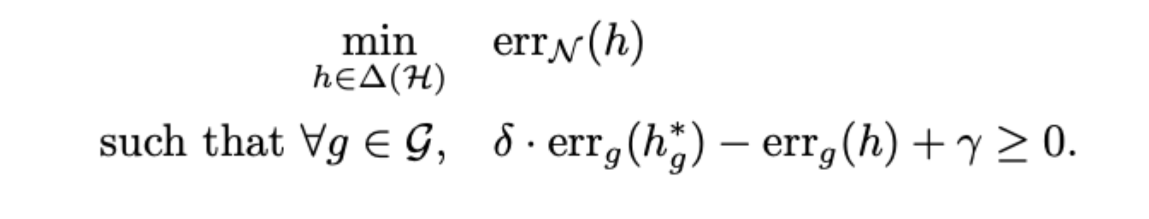
- Akin to groups being completely unknown.
- For this case, we devise the *Proportional Fairness (PF)* classifier: the randomized classifier h_{PF} that maximizes the sum over data points i, $\sum_i \log(u_i(h))$, where $u_i(h)$ is the expected accuracy of a randomized classifier h on the data point i.
- PF achieves the following guarantee:: For any group g with a perfect classifier h_g^*,

  the accuracy $u_g(h_{PF})$ of h_{PF} on g satisfies $u_g(h_{PF}) \geq \frac{|g|}{N}$,

  Where $|g|$ is the size of the group g, and N is the size of the entire data set.
- In general, one cannot do better than the bound given above.
- One drawback of PF is that the theoretical guarantee is *proportionally lower for smaller groups* (although in practice, the accuracy achieved by PF on small groups is much better than what is given by these bounds).

Linear separable groups

- For linear separable groups we can do much better.
- For this case, we formulate the BeFair problem:

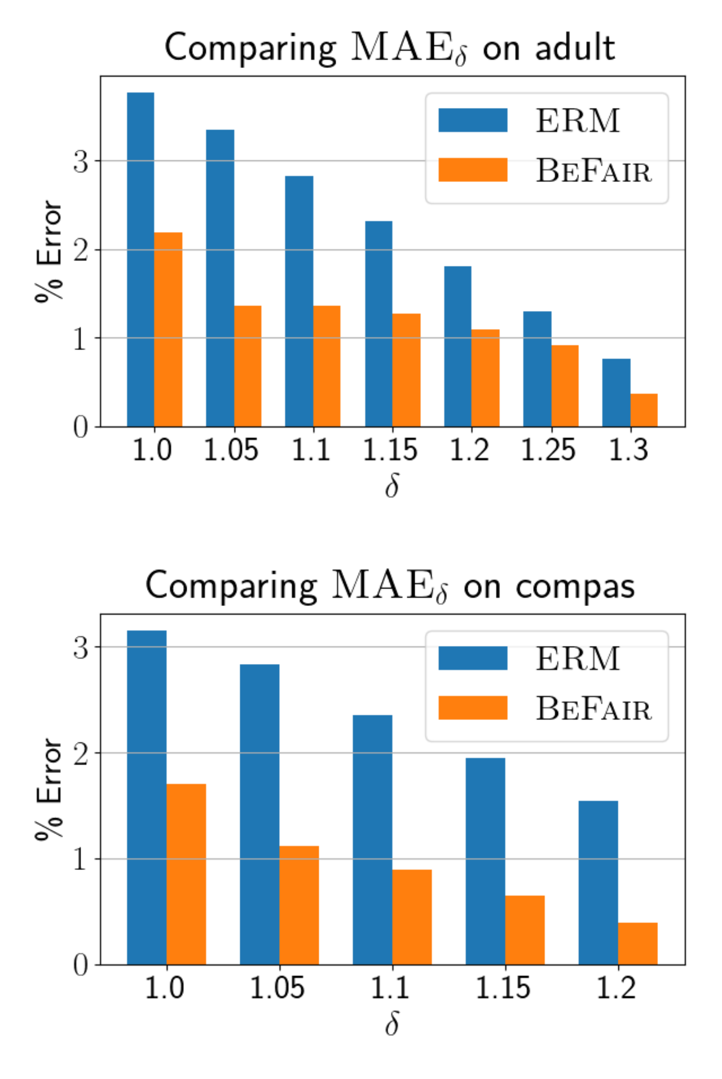


$$\begin{aligned} \min_{h \in \Delta(\mathcal{H})} \quad & \mathrm{err}_{\mathcal{N}}(h) \\ \text{such that } \forall g \in \mathcal{G}, \quad & \delta \cdot \mathrm{err}_g(h_g^*) - \mathrm{err}_g(h) + \gamma \geq 0. \end{aligned}$$

- In other words, the goal is to approximate the expected accuracy on every group, linearly with respect to the accuracy of the optimal classifier for that group.
- Note that the guarantee is the *same irrespective of the size of the group*.
- We solve the above problem using novel convex relaxation techniques in an overall adversarial optimization framework.

Experiments

- We work with a few data sets, one of which is adult, the Adult income dataset from the UCI Machine Learning Repository.
- We show that our algorithms work well in practice.
- For example, BeFair is able to achieve a performance that is a close approximation of the best possible on these groups.



Conclusion

- We studied best-effort guarantees of fairness: on each group, achieving an accuracy that is as close as possible to the optimal accuracy on that group alone.
- We devised classification methods that achieve such guarantees.
- Future work: Many interesting questions for future work:
  - Looking at more general non-linear classes of groups for BeFair.
  - How to think about best-effort fairness for deterministic classifiers?

THANK YOU!