A Sandbox Tool to Bias(Stress)-Test Fairness Algorithms

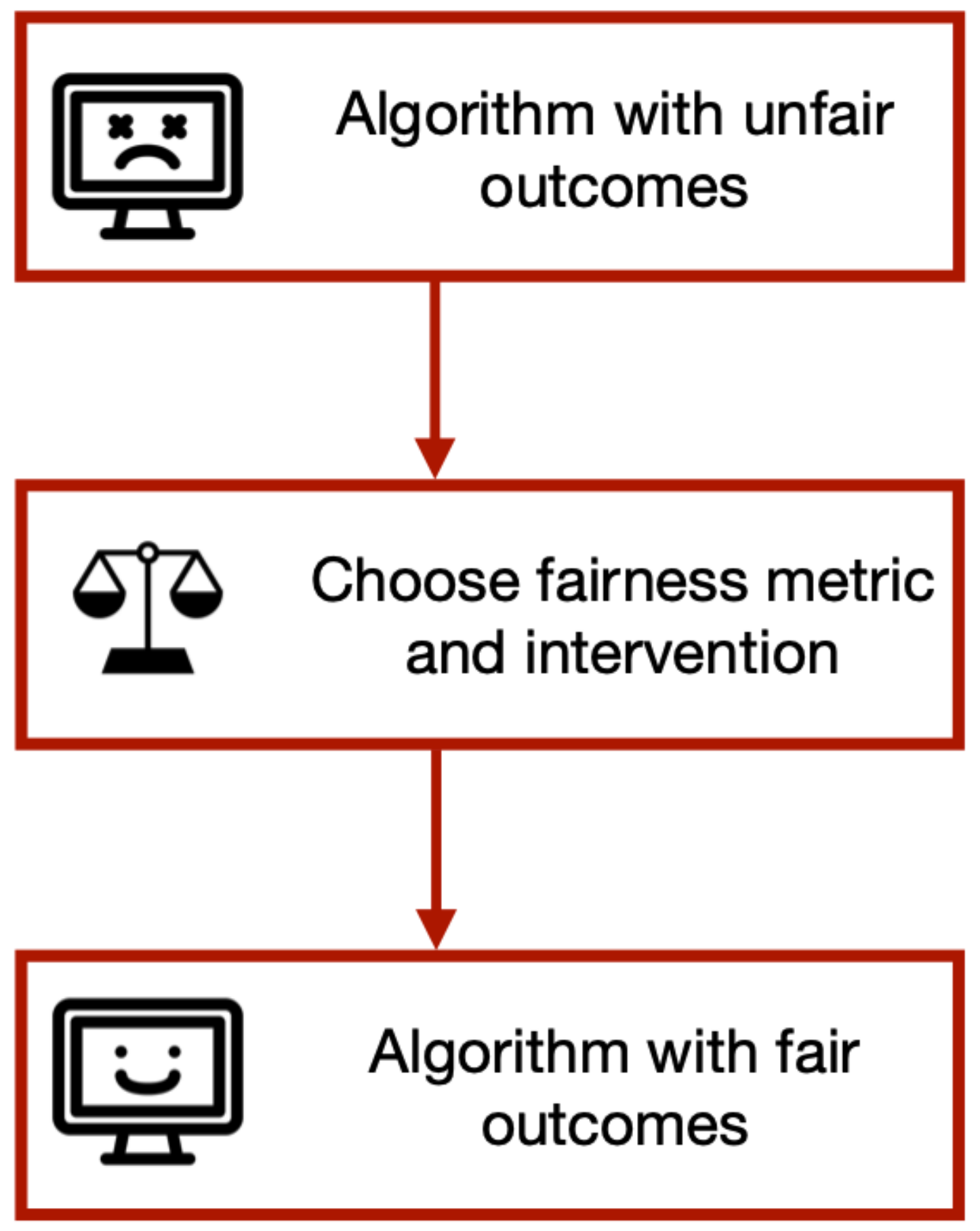
Nil-Jana Akpinar [1] Manish Nagireddy [1] Logan Stapelton [2] Hao-Fei Cheng [2] Haiyi Zhu [1] Steven Wu [1] Hoda Heidari [1]

[1]Carnegie Mellon University [2]University of Minnesota

Motivation & Contributions

Motivation

- Figure 1 illustrates a **common fair ML pipeline** that selects an off-the-shelf fairness algorithm.
- **Problem:** Most fairness-enhancing algorithms are agnostic to the *source* of unfairness.
- Blind application may hide the real problem by ensuring narrowly defined notions of fairness.



Algorithm with unfair outcomes

Choose fairness metric and intervention

Algorithm with fair outcomes

Figure 1. Common fair ML narrative

Contributions: The Sandbox Tool

- We offer a **simulation framework** for examining fairness interventions in the presence of counterfactually injected biases (e.g. representation bias, measurement bias, omitted variable bias, model validity discrepancies).
- Allows us to test whether a given remedy can alleviate a particular type of bias by comparing results before and after bias injection.

Intended Use of the Tool

- **Research settings:** to explore relationship between bias and unfairness, and shape informed hypotheses,
- **Educational settings:** to demonstrate the nuanced sources of unfairness, and grasp the limitations of fairness-enhancing algorithms,
- **Practical settings:** to explore the potential effect of various algorithmic interventions in real-world applications *if and only if* the bias pattern is well-understood.

Description of the Sandbox

The Sandbox framework is comprised of six modular steps with room for customization at each stage.

1. **Choice of Data:** Synthetic data generation, upload data set, or choose pre-loaded data set.
2. **Bias Injection:** Select from different types of bias (e.g. representation bias, measurement bias, omitted variable bias), and select sub-group to inject bias into.
3. **Model Class Selection:** Select model type to fit on data set (e.g. logistic regression, scikit-learn classifiers)
4. **Fairness Intervention:** Choose bias mitigation intervention (pre-, in-, or post-processing) implemented with Fairlearn.
5. **Evaluation Metrics:** Provide a list of evaluation metrics to output (e.g. accuracy, various fairness metrics, fidelity).
6. **Visualization:** Outputs visualizations of effectiveness of the fairness intervening at mitigating bias and improving accuracy.

Case study: Can Fairness Improve Accuracy?

For demonstration, we use the sandbox tool to empirically explore the performance of a known theoretical result

> **Blum and Stangl [1]:** *In specific settings, Equalized Odds (i.e. TPR and FPR equal across groups) constrained empirical risk minimization on data with under-representation bias (i.e. remove rows with positive labels from minority group) can recover the Bayes optimal classifier on the true data.*

The **sandbox tool** can help to

- Give a sense of how fast the result kicks in with a finite sample,
- Assess the effectiveness in a specific data generation and hypothesis class setting,
- Understand the importance of different assumptions for the result.

Results: Case Study & Exploration

Case Study

- Generate synthetic data and simulate the setting assumed in the theory: 3 features, group-dependent linear Bayes optimal classifiers, label noise, 7 parameter logistic regression.
- Even under these very favorable conditions, a lot of data is required for the result to kick in (Figure 2).
- Many practical applications fall into the range of small data sets and moderate underrepresenation bias in which the intervention was unsuccessful.

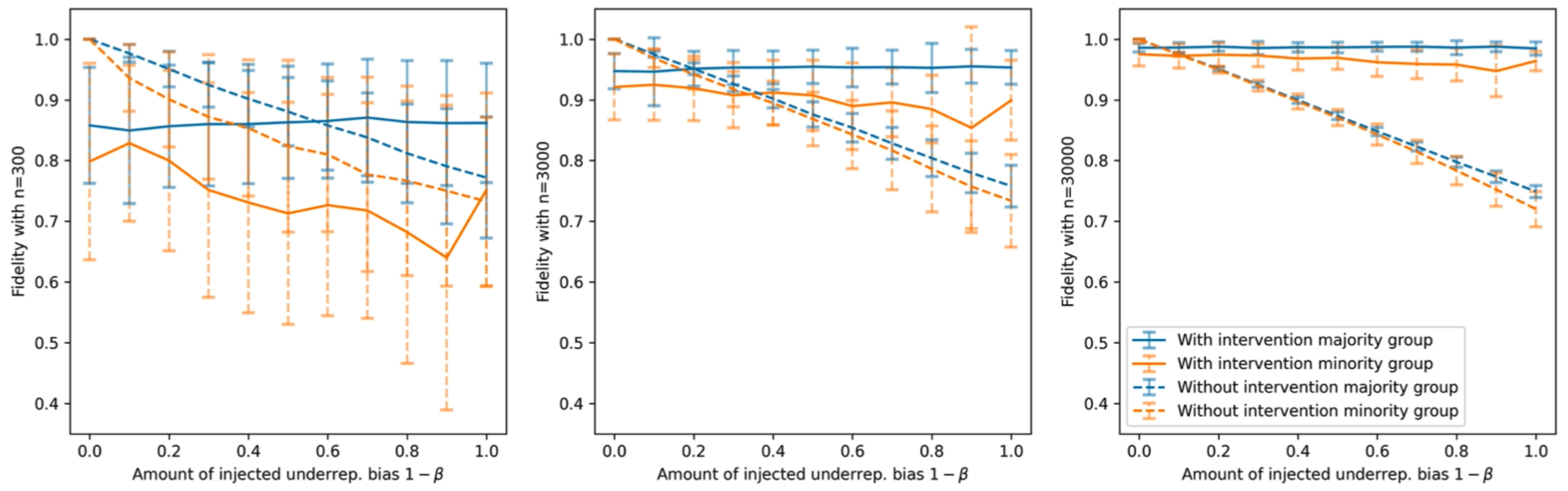


Figure 2. Test set fidelity between Bayes optimal classifier and models trained on biased data with and without fairness intervention. Test set unbiased.

Exploration

- When loosening the assumptions or injecting different types of bias, the Equalized Odds intervention struggles to recover the Bayes optimal model
- See Figure 3 for **Difference in Base Rates** and **Label Bias** settings.
- In both settings, the model may seem fair but the Bayes optimal model cannot fulfill Equalized Odds analytically.

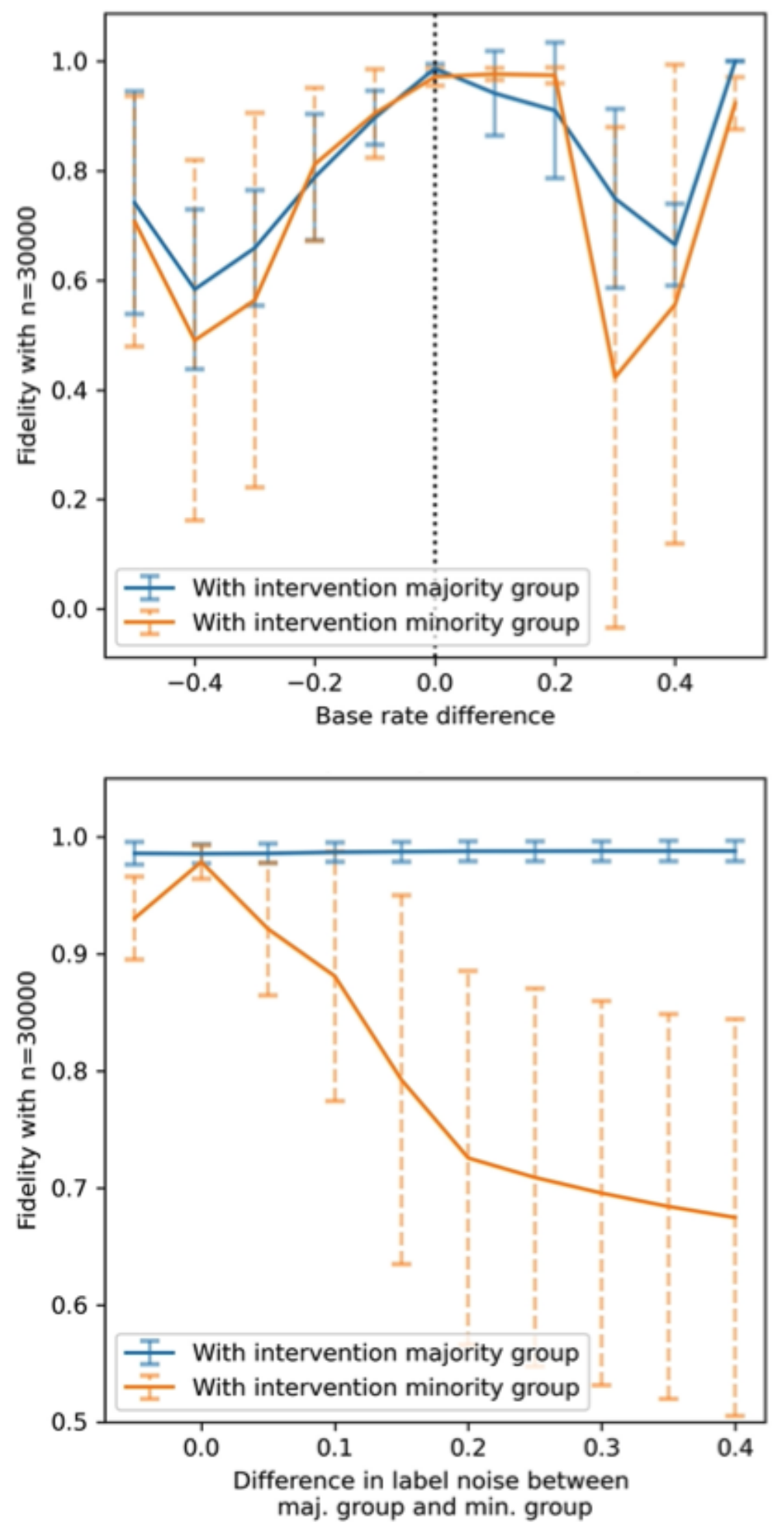


Figure 3. Test set fidelity when loosening assumptions.

References

[1] Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *Proceedings of the 2020 Symposium on the Foundations of Responsible Computing* (FORC), 2020.