Making Decisions that Reduce Discriminatory Impact

Matt J. Kusner[1,2] Chris Russell[1,3] Joshua R. Loftus[4] Ricardo Silva[1,5]

[1]Alan Turing Institute [2]University of Oxford [3]University of Surrey [4]New York University [5]University College London

Focal points

- The **discriminatory impact** problem
- Fair interventions under **interference**

Summary

Defining impact: an event caused jointly by the decisions under our control *and* other real world factors. In particular, *decisions about one individual can impact another individual*. In a fairness setting we want to ensure the intervention does not have an unfair impact related to sensitive attribute(s) **A** like gender and/or race. (See also [1, 2, 3])

Fair predictions/decisions do not imply fair impacts, since other downstream factors can make the impact unfair (possibly to different individuals than the subjects of the original prediction/decision). We demonstrate a focus on impact in two ways:

- Formally modeling the intervention as a separate variable **Z** from the (predicted) outcome **Y** in a structural causal model (SCM). Our goal is an optimal, fair intervention.
- Modeling interference: the intervention targeted for one unit may causally impact other units [4, 5].

Fair interventions: we define fairness and design fair interventions in the context of an SCM relating the observed data, predicted outcome, and planned intervention. Most closely related to the counterfactual fairness framework of [1].

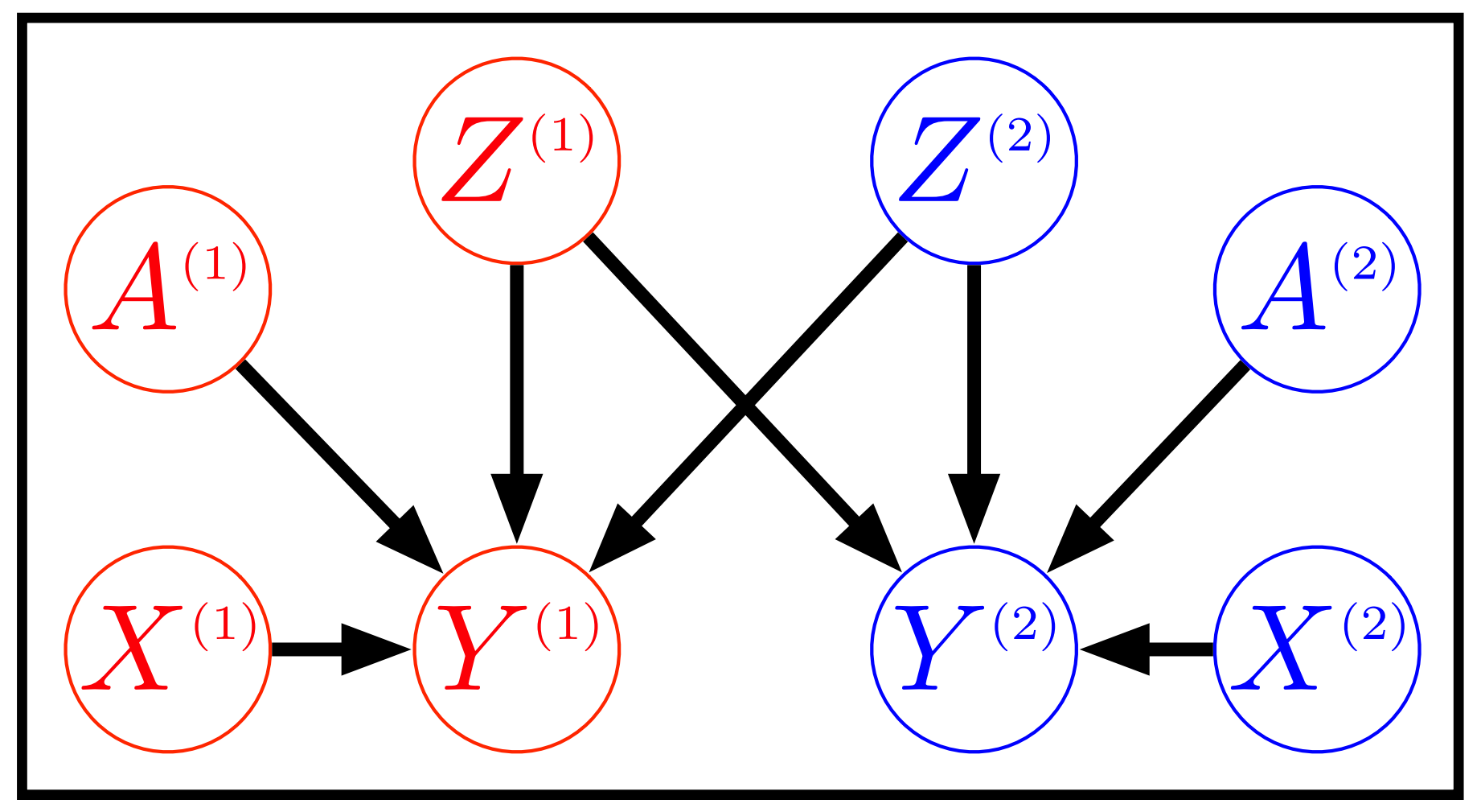

Figure 1: An example causal model (SCM) illustrating interference. Sensitive attribute **A**, outcome **Y**, other predictors **X**, intervention **Z**

Counterfactual fairness

We try to understand fairness using a causal model (SCM) as follows:

1. Change a sensitive attribute value from a to a'
2. Propagate this change through the SCM to descendants of A
3. Use resulting counterfactual (hypothetical) values to audit fairness

Counterfactual privilege

An asymmetric fairness constraint assuming larger values of **Y** are desirable

$$\begin{aligned} c_{ia'} :=& \mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = {}^{(i)}] \\ &- \mathbb{E}[Y^{(i)}(a', \mathbf{z}) \quad \mid A^{(i)} = a^{(i)}, X^{(i)} = {}^{(i)}] < \tau \end{aligned}$$

In practice will only be active for privileged actual values a, capturing the intuition that the intervention should not help units who are already doing well because of their privilege sensitive attribute.

Simulation results

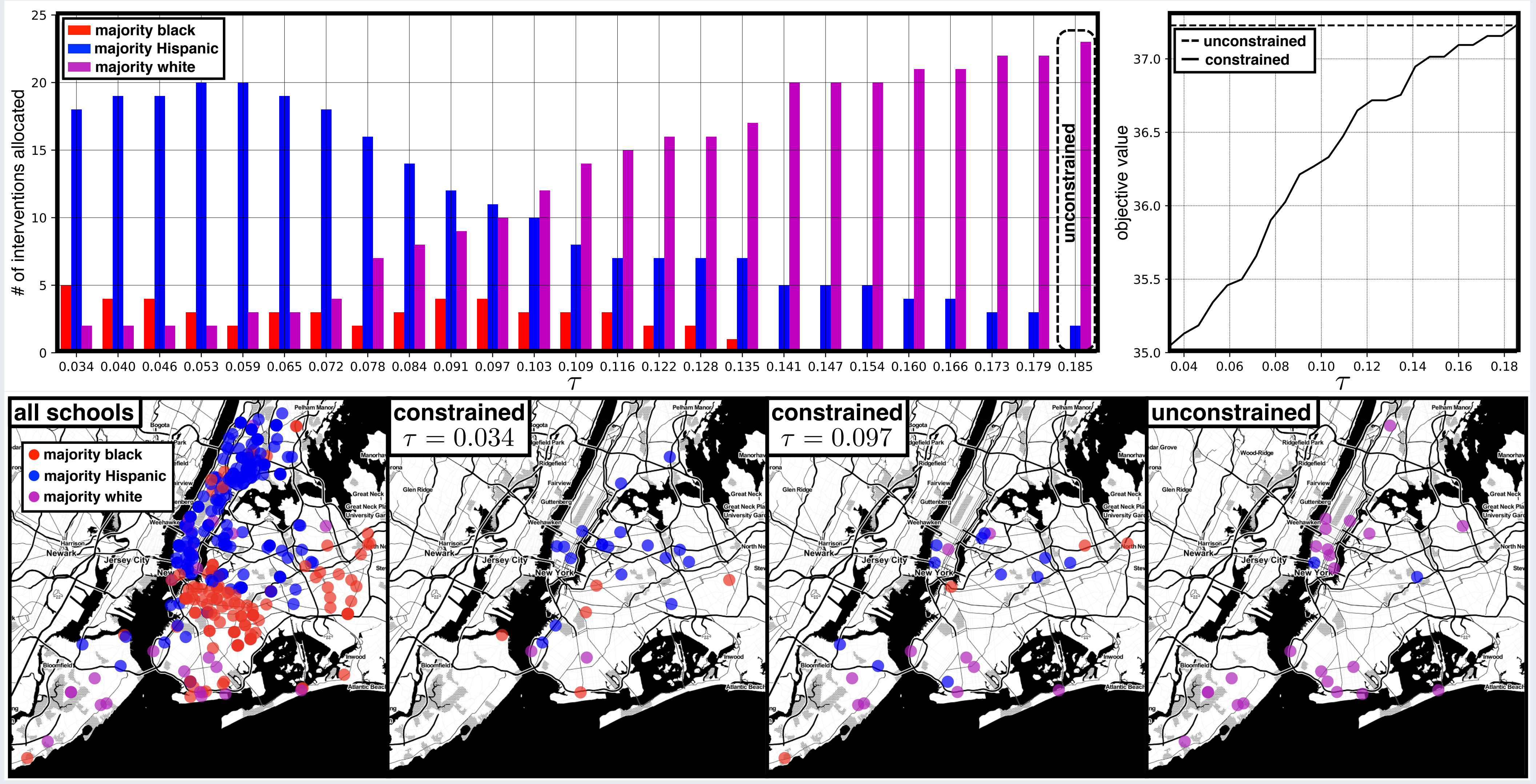

Figure 2: School intervention simulation results for different values of counterfactual privilege constraint τ

Simulation example setup

- Budget B to pay for new calculus classes
- Intervention: $\mathbf{Z}^{(i)} = 1$ if school i receives funding for a class and 0 otherwise
- Outcome: $\mathbf{Y}^{(i)}$ percent of students at school i taking the SAT (planning to go to college)
- Protected attribute: $\mathbf{A}^{(i)}$ encodes whether school i is majority black, Hispanic, or white
- Interference: students at school i may be able to take a calculus class at *nearby schools*

Algorithm

We use a mixed-integer linear program (MILP) to solve

$$\begin{aligned} \max_{\mathbf{z} \in \{0,1\}^n} & \sum_{i=1}^{n} \mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = {}^{(i)}] \\ s.t., & \sum_{i=1}^{n} z^{(i)} \leq B \\ & c_{ia'} \leq \tau \quad \forall a' \in \mathcal{A},\ i \in \{1, \ldots, n\}, \end{aligned}$$

where $\mathcal{A}$ is the domain of A and $\tau \geq 0$.

Conclusion

We made specific choices to concretely illustrate the main ideas, but the overall approach is much more broadly applicable. We must attempt to understand fairness in the context of the real world, not just the training data and prediction task.

References

[1] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva.
Counterfactual fairness.
In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[2] Lydia Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt.
Delayed impact of fair machine learning.
In *International Conference on Machine Learning*, pages 3156–3164, 2018.

[3] Ben Green and Yiling Chen.
Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments.
In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99. ACM, 2019.

[4] Elizabeth L Ogburn, Tyler J VanderWeele, et al.
Causal diagrams for interference.
Statistical science, 29(4):559–578, 2014.

[5] Peter M Aronow, Cyrus Samii, et al.
Estimating average causal effects under general interference, with application to a social network experiment.
The Annals of Applied Statistics, 11(4):1912–1947, 2017.

Acknowledgements

Supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. CR acknowledges additional support under the EPSRC Platform Grant EP/P022529/1

Contact Information

loftus@nyu.edu - http://joshualoftus.com