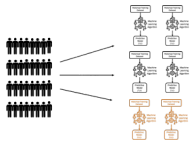
Fair Machine Learning Under Partial Compliance

Jessica Dai[1], Sina Fazelpour[2], Zachary C. Lipton[2]

[1]Brown University, [2]Carnegie Mellon University

MOTIVATION

While most work in fair machine learning focuses on the outputs of a single algorithm in isolation, many real-world scenarios involve *multiple competing decisionmakers.*



Many (competing) decisionmakers
Only some care about fairness
Individuals choose where to apply

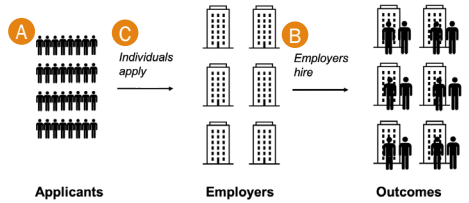
Partial compliance occurs when only some decisionmakers care about fairness. In a competitive marketplace, partial compliance means that the output of individual decision-makers cannot be considered in a vacuum:

1. *Interactive effects.* The decisions made by one institution will affect the inputs (candidates) seen by other institutions in the future.
2. *Strategic behavior*. Individual candidates may alter their application strategy, meaning each decisionmaker may see a different distribution of candidates.

Core question: What are the implications of partial compliance in light of the dynamic interactions that may emerge between individuals and institutions?

SIMULATION SETUP

We use the labor market as a toy model for our simulations.



A. Applicant population: all individuals are described by exactly two features: group membership, and score (representing some notion of qualification).

B. Hiring policies: all employers are either non-compliant or compliant; all compliant employers in a single simulation use the same policy.

1. **Non-compliant:** hire solely based on score
2. **Compliant:** satisfy some version of demographic parity.
   Global parity: satisfy DP w/r/t global demographics
   Local parity: satisfy DP w/r/t current applications

C. Application strategies: each group has an application strategy reflecting preference for a compliant vs non-compliant employer.

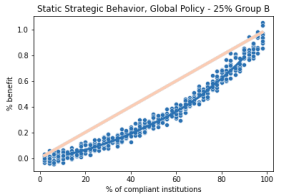
1. **Static strategy** (limited knowledge): each group has a slight static preference for the employer type at which they are more likely to be hired
2. **Adaptive strategy** (access to new information): at each timestep, each group updates their preferences based on results from the previous round.

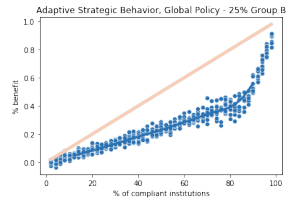
RESULTS

1. Sublinear gain: *k*% compliance does not bring *k*% benefit.

"benefit": demographic parity $\frac{P(hired\ |B)}{P(hired\ |A)}$ (scaled by baseline DP at 0% compliant).

Both graphs: global parity policy. Left: static applicant strategy
Right: adaptive applicant strategy

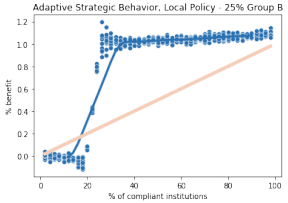


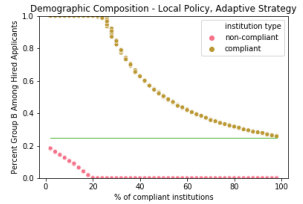


2. The emergence of segregation under specific parameter settings

Both graphs: adaptive applicant strategy and local parity policy.

Left: % compliant vs % benefit
Right: demographic composition (% of hired employees belonging to group B)

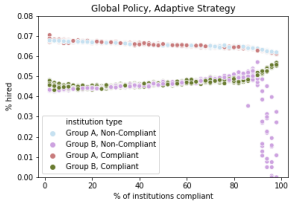




3. The difficulty of auditing for non-compliance

Suppose we ask **"what % of applicants from Group X do you hire?"** to each employer type.

Under global parity policy and adaptive applicant strategy, compliant and non-compliant employers are indistinguishable!



All experiments run with total 50 employers, varying the number of compliant institutions from 0 to 50. Statistics calculated based on ***post-equilibrium*** *timesteps.*

Key takeaways: partial compliance (& dynamic behavior) can drastically impact downstream effects of fair policies; as a result, the evaluation of "fair algorithms" must consider the wider environment of deployment.

https://arxiv.org/abs/2011.03654 * jessica.dai@alumni.brown.edu