# Fair Prediction with Endogenous Behavior

Changhwa Lee (Speaker), Christopher Jung, Sampath Kannan, Mallesh Pai, Aaron Roth, and Rakesh Vohra

## Algorithmic Fairness

- Debate: What kinds of fairness measures for classification are desirable?
- Common way to think: given the data, propose a fairness measure and an algorithm that achieves it, and test with the data.
- We argue: taking agents' endogenous behavior into account is important, in the context of criminal justice system.

# What we do

**Model**

Agents from different groups decide whether to commit a crime or not, by comparing payoffs of crime and probability of being classified as guilty.

Judge designs a classification rule to minimize the average crime rate.

**Crime Minimizing Classification**

Crime-minimizing classification maximizes disincentive to commit a crime.

# Properties of Crime Minimizing Classification

**Crime Minimizing Classification**

1. Crime-minimizing classification *only* cares about giving the right incentive to induce the right behavior.

2. Fair in equalizing: false positive rates, false negative rates and disincentives.

3. Incompatible with: equalizing posterior risk thresholds, equalizing positive / negative parity rates.

# Robustness

That the incentive is the *only* thing that matters is robust:

- Agents may have different costs and rewards for crime.
- Agents may not behave perfectly rationally and pick (e.g.) a random action with some probability $q_i$
- Signals may be observed at different rates across the groups.
- If the signal distributions differ by group, for a large class of signal structures, equalizing disincentives $\Delta_g$ is the only fairness notion that is compatible with the crime-minimizing classification.

**Takeaway:** Incentives matter!