Experimental Evaluation of Computer-Assisted Human Decision Making

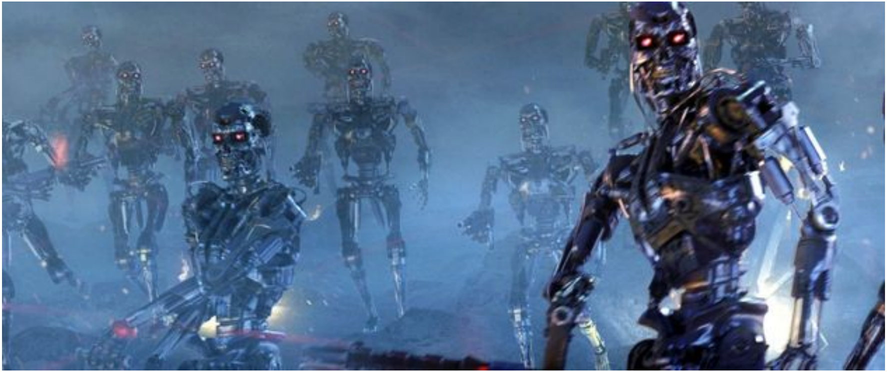
Kosuke Imai

Harvard University

Quantitative Social Science Colloquium
Princeton University
February 28, 2020

Joint work with Zhichao Jiang (UMass. Amherst)
Jim Greiner and Ryan Halen (Harvard Law School)

Rise of the Machines



- Statistics, machine learning, artificial intelligence in our daily lives
- Nothing new but accelerated due to technological advances
- Examples: factory assembly lines, ATM, home appliances, autonomous cars and drones, games (Chess, Go, Shogi), ...

Motivation

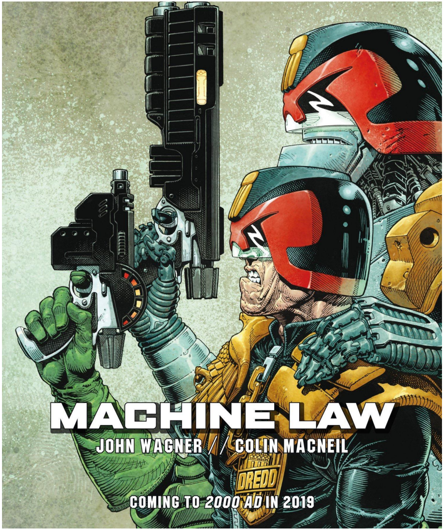
- But, humans still make many consequential decisions
  - this is true even when human decisions can be suboptimal
  - we may want to hold *someone*, rather than *something*, accountable
- Computer-assisted human decision making
  - humans make decisions with the aid of machine recommendations
  - routine decisions made by individuals in daily lives
  - consequential decisions made by judges, doctors, etc.
- How do machine recommendations influence human decisions?
  - Do they help human decision-makers achieve a goal?
  - Do they help humans improve the fairness of their decisions?
- Many have studied the accuracy and fairness of machine recommendations rather than their impacts on human decisions
- We develop a set of statistical methodology for experimentally evaluating computer-assisted human decision making

Application: Pretrial Risk Assessment Instrument

- Machine recommendations often used in US criminal justice system
- At the first appearance hearing, judges primarily make two decisions
  1. whether to release an arrestee pending disposition of criminal charges
  2. what conditions (e.g., bail and monitoring) to impose if released
- Goal: avoid predispositional incarceration as much as possible if safe
- Judges are required to consider two risk factors along with others
  1. arrestee may fail to appear in court (FTA)
  2. arrestee may engage in new criminal activity (NCA) if released
- PRAI as a machine recommendation to judges
  - classifying arrestees according to FTA and NCA risks
  - derived from an application of a machine learning algorithm or a statistical model to a training data set based on past observations
- Controversy over the potential racial bias of COMPAS score
  - Propublica's analysis and Northpointe's rebuttal
  - Almost all existing work focus on the accuracy and fairness of PRAI

But, Machines Do Not Make Judicial Decisions for Us

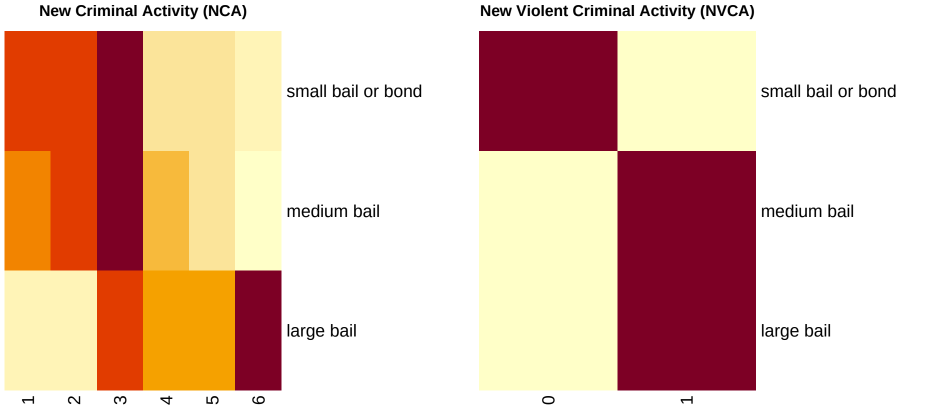
Well, at least not yet...



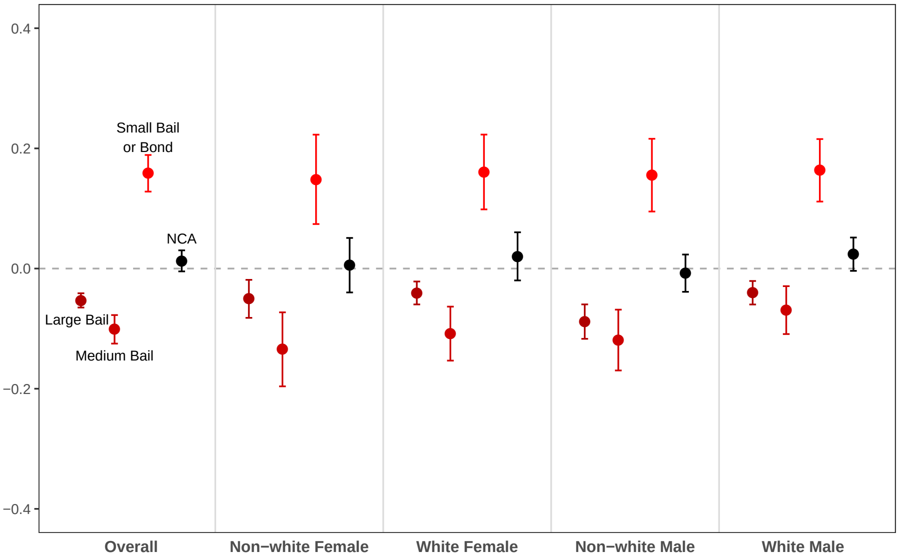
A Field Experiment for Evaluating a PRAI

- A Midwestern county
- PRAI
  - based on criminal history (prior convictions and FTA) and age
  - two separate ordinal risk scores for FTA and NCA
  - one binary risk score for new violent criminal activity (NVCA)
- Judges have other information about an arrestee
  - affidavit by a police officer about the arrest
  - defense attorney may inform about the arrestee's connections to the community (e.g., family, employment)
  - assistant district attorney may provide additional information
- Field experiment
  - clerk assigns case numbers sequentially as cases enter the system
  - PRAI is calculated for each case using a computer system
  - if the first digit of case number is even, PRAI is given to the judge
- Prior work
  - mostly observational studies or hypothetical survey experiments
  - only exception: The 1981 – 82 Philadelphia bail experiment

A (Somewhat Empirically Informed) Synthetic Data Set



Intention-to-Treat Analysis of PRAI Provision



The Setup of the Proposed Methodology

- Notation:
  - $i = 1, 2, \ldots, n$: cases
  - Z_i: whether PRAI is presented to the judge ($Z_i = 1$) or not ($Z_i = 0$)
  - D_i: judge's binary decision to release ($D_i = 1$) or detain ($D_i = 0$)
  - Y_i: binary outcome (NCA, FTA, or NVCA)
  - X_i: observed (by researchers) pre-treatment covariates

- Potential outcomes:
  - $D_i(z)$: potential value of the release decision when $Z_i = z$
  - $Y_i(z, d)$: potential outcome when $Z_i = z$ and $D_i = d$
  - Relationship to observed data: $D_i = D_i(Z_i)$ and $Y_i = Y_i(Z_i, D_i(Z_i))$
  - No interference across cases: can analyze the first arrest cases only

- Assumptions maintained throughout our analysis:
  1. Randomized treatment assignment: $\{D_i(z), Y_i(z,d), \mathrm{X}_i\} \perp\!\!\!\perp Z_i$
  2. Exclusion restriction: $Y_i(z, d) \ = \ Y_i(d)$
  3. Monotonicity: $Y_i(0) \leq Y_i(1)$

Causal Quantities of Interest

- Principal stratification (Frangakis and Rubin 2002)
  - $(Y_i(1), Y_i(0)) = (1, 0)$: preventable cases
  - $(Y_i(1), Y_i(0)) = (1, 1)$: risky cases
  - $(Y_i(1), Y_i(0)) = (0, 0)$: safe cases
  - ~~$(Y_i(1), Y_i(0)) = (0, 1)$~~: eliminated by monotonicity

- Average causal effects of PRAI on judge's decisions:

$$\begin{aligned} \text{ACEp} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 1, Y_i(0) = 0\}, \\ \text{ACEr} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 1, Y_i(0) = 1\}, \\ \text{ACEs} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 0, Y_i(0) = 0\}. \end{aligned}$$

- If PRAI is helpful, we should have ACEp < 0 and ACEs > 0
- The desirable sign of ACEr depends on various factors

Partial Identification

- Under the assumptions of randomization, exclusion restriction, and monotonicity, we have

$$\begin{aligned}
\text{ACEp} &= \frac{\Pr(D_i = 1, Y_i = 1 \mid Z_i = 1) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 1)}{\Pr\{Y_i(1) = 1\} - \Pr\{Y_i(0) = 1\}} \\
&\quad - \frac{\Pr(D_i = 1, Y_i = 1 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 0)}{\Pr\{Y_i(1) = 1\} - \Pr\{Y_i(0) = 1\}}, \\
\text{ACEr} &= \frac{\Pr(D_i = 0, Y_i = 1 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 1 \mid Z_i = 1)}{\Pr\{Y_i(0) = 1\}}, \\
\text{ACEs} &= \frac{\Pr(D_i = 0, Y_i = 0 \mid Z_i = 1) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 0)}{1 - \Pr\{Y_i(1) = 1\}}.
\end{aligned}$$

- The signs are identified since $Y_i(0) \leq Y_i(1)$
- The bounds can be derived using the law of iterated expectation

$$\begin{aligned}
\Pr\{Y_i(d) = 1\} &= \Pr\{Y_i = 1 \mid D_i = d\} \Pr(D_i = d) \\
&\quad + \Pr\{Y_i(d) = 1 \mid D_i = 1 - d\} \Pr(D_i = 1 - d)
\end{aligned}$$

for $d = 0, 1$

Point Identification under Unconfoundedness

- Unconfoundedness:

$$Y_i(d) \perp\!\!\!\perp D_i \mid \mathrm{X}_i, Z_i = z$$

for $z = 0, 1$ and all d.

- Violated if judges base their decision on additional information they have about arrestees $\rightsquigarrow$ sensitivity analysis

- Principal scores (Ding and Lu 2017)

$$\begin{aligned} e_P(\mathrm{x}) &= \Pr\{Y_i(1) = 1, Y_i(0) = 0 \mid \mathrm{X}_i = \mathrm{x}\} \\ e_R(\mathrm{x}) &= \Pr\{Y_i(1) = 1, Y_i(0) = 1 \mid \mathrm{X}_i = \mathrm{x}\} \\ e_S(\mathrm{x}) &= \Pr\{Y_i(1) = 0, Y_i(0) = 0 \mid \mathrm{X}_i = \mathrm{x}\} \end{aligned}$$

Identification Results

Under the assumptions of randomization, monotonicity, exclusion restriction, and unconfoundedness, we can identify causal effects as

$$
\begin{aligned}
\text{ACEp} &= \mathbb{E}\{w_P(\mathsf{X}_i) D_i \mid Z_i = 1\} - \mathbb{E}\{w_P(\mathsf{X}_i) D_i \mid Z_i = 0\}, \\
\text{ACEr} &= \mathbb{E}\{w_R(\mathsf{X}_i) D_i \mid Z_i = 1\} - \mathbb{E}\{w_R(\mathsf{X}_i) D_i \mid Z_i = 0\}, \\
\text{ACEs} &= \mathbb{E}\{w_S(\mathsf{X}_i) D_i \mid Z_i = 1\} - \mathbb{E}\{w_S(\mathsf{X}_i) D_i \mid Z_i = 0\},
\end{aligned}
$$

where

$$
w_P(\mathsf{x}) = \frac{e_P(\mathsf{x})}{\mathbb{E}\{e_P(\mathsf{X}_i)\}}, \quad w_R(\mathsf{x}) = \frac{e_R(\mathsf{x})}{\mathbb{E}\{e_R(\mathsf{X}_i)\}}, \quad w_S(\mathsf{x}) = \frac{e_S(\mathsf{x})}{\mathbb{E}\{e_S(\mathsf{X}_i)\}}.
$$

and

$$
\begin{aligned}
e_P(\mathsf{x}) &= \Pr\{Y_i = 1 \mid D_i = 1, \mathsf{X}_i = \mathsf{x}\} - \Pr\{Y_i = 1 \mid D_i = 0, \mathsf{X}_i = \mathsf{x}\}, \\
e_R(\mathsf{x}) &= \Pr\{Y_i = 1 \mid D_i = 0, \mathsf{X}_i = \mathsf{x}\}, \\
e_S(\mathsf{x}) &= \Pr\{Y_i = 0 \mid D_i = 1, \mathsf{X}_i = \mathsf{x}\}.
\end{aligned}
$$

Extension to Ordinal Decision

- Judge's decision is typically ordinal (e.g., bail amount)
- $D_i = 0, 1, \ldots, k$: a bail of *decreasing* amount
- Monotonicity: $Y_i(d_1) \ \leq \ Y_i(d_2) \quad \text{for} \quad d_1 \leq d_2$
- Principal strata based on an ordinal measure of safety

$$R_i \ = \ \begin{cases} \max\{d : Y_i(d) = 0\} & \text{if } \ Y_i(0) = 0 \\ -1 & \text{if } \ Y_i(0) = 1 \end{cases}$$

- Least amount of bail that keeps an arrestee from committing NCA
- Example with $k = 2$: risky cases ($R_i = -1$), preventable cases (high risk $R_i = 0$; low risk $R_i = 1$), safe cases ($R_i = 2$)
- Causal quantities of interest:

$$\text{ACEp}(r) \ = \ \Pr\{D_i(1) \leq r \mid R_i = r\} - \Pr\{D_i(0) \leq r \mid R_i = r\}$$

for $r = 0, 1, \ldots, k-1$
  - reduction in the proportion of NCA attributable to the PRAI
  - $n\mathbb{E}\{\text{ACEp}(R_i)\}$: expected number of NCAs prevented

Identification for Ordinal Decision

- Safe cases ($R_i = k$) $\rightsquigarrow$ should be released

$$\text{ACEs} = \Pr\{D_i(1) = k \mid R_i = k\} - \Pr\{D_i(0) = k \mid R_i = k\}.$$

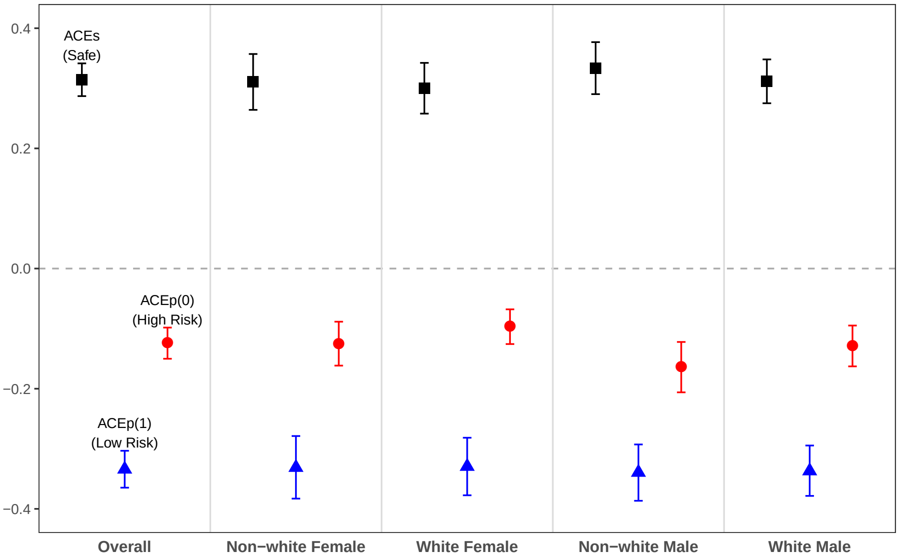
- Identification under unconfoundedness:

$$\begin{aligned} \text{ACEp}(r) &= \mathbb{E}\{w_r(\mathsf{X}_i)1(D_i \leq r) \mid Z_i = 1\} \\ &\quad - \mathbb{E}\{w_r(\mathsf{X}_i)1(D_i \leq r) \mid Z_i = 0\}, \\ \text{ACEs} &= \mathbb{E}\{w_k(\mathsf{X}_i)1(D_i = k) \mid Z_i = 1\} \\ &\quad - \mathbb{E}\{w_k(\mathsf{X}_i)1(D_i = k) \mid Z_i = 0\}, \end{aligned}$$

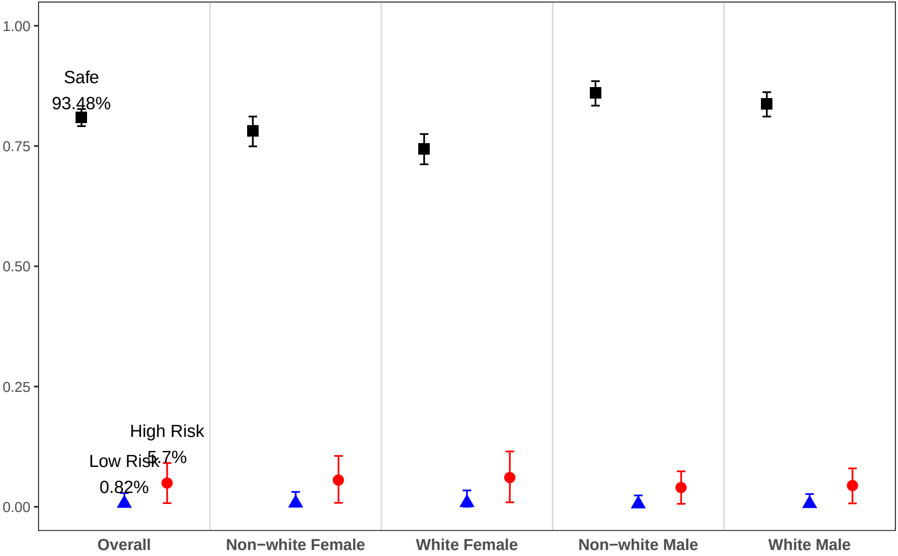
where

$$\begin{aligned} w_r(\mathsf{x}) &= e_r(\mathsf{x})/\mathbb{E}\{e_r(\mathsf{X}_i)\}, \\ e_r(\mathsf{x}) &= \Pr(R_i = r \mid \mathsf{X}_i = \mathsf{x}) \\ &= \Pr\{Y_i = 1 \mid D_i = r+1, \mathsf{X}_i = \mathsf{x}\} \\ &\quad - \Pr\{Y_i = 1 \mid D_i = r, \mathsf{X}_i = \mathsf{x}\} \text{ for } r = 0, 1, \ldots, k-1, \\ e_k(\mathsf{x}) &= \Pr\{Y_i(k) = 0 \mid \mathsf{X}_i = \mathsf{x}\} = \Pr\{Y_i = 0 \mid D_i = k, \mathsf{X}_i = \mathsf{x}\}. \end{aligned}$$

Estimated Average Causal Effects



Estimated Proportion of Principal Strata



Sensitivity Analysis

- Judges may use additional information when making decisions
- Bounds: avoid the unconfoundedness assumption
- Sensitivity analysis: How robust are one's empirical results to the potential violation of the key assumption?
- Ordinal probit models for $D_i(z)$ and R_i with latent variables

$$
\begin{aligned}
D_i^*(z) &= \beta z + \mathrm{X}_i^\top \gamma + \epsilon_{i1}, \\
R_i^* &= \boldsymbol{X}_i^\top \alpha + \epsilon_{i2},
\end{aligned}
$$

where $\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$

- Identified under unconfoundedness (i.e., $\rho = 0$)
- R_i is not observable but $R_i \leq r - 1 \Longleftrightarrow Y_i(r) = 1$

$$
\Pr\{Y(r) = 1\} \;=\; \Pr\{R_i^* \leq \delta_r\} = \Pr(\delta_r - \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2} > 0).
$$

where δ_r is the rth threshold for R_i

Principal Fairness

- Literature focuses on the fairness of machine-recommendations/PRAI
- We focus on the fairness of human decision
- Problems with the existing definitions and methods:
  1. protected attributes should not be used as inputs
     $\rightsquigarrow$ may still depend on these attributes through other variables
  2. equality of classification accuracy between different groups
     $\rightsquigarrow$ censoring may bias the results
  3. counterfactual fairness: what if one belongs to a different group
     $\rightsquigarrow$ many attributes cannot be manipulated

- Principal fairness: decision should not (statistically) depend on a protected attribute S_i within a principal strata

$$D_i \perp\!\!\!\perp S_i \mid R_i = r \quad \text{for all } r \in \{-1, 0, 1, \ldots, k\}$$

Measuring and Estimating the Degree of Fairness

- How fair are the judges' decisions?

$$\begin{aligned}\Delta_r(z) \;=\; \max_{s,s'} \big| & \Pr\{D_i(z) \leq r \mid S_i = s, R_i = r\} \\ & - \;\Pr\{D_i(z) \leq r \mid S_i = s' R_i = r\}\big|\end{aligned}$$

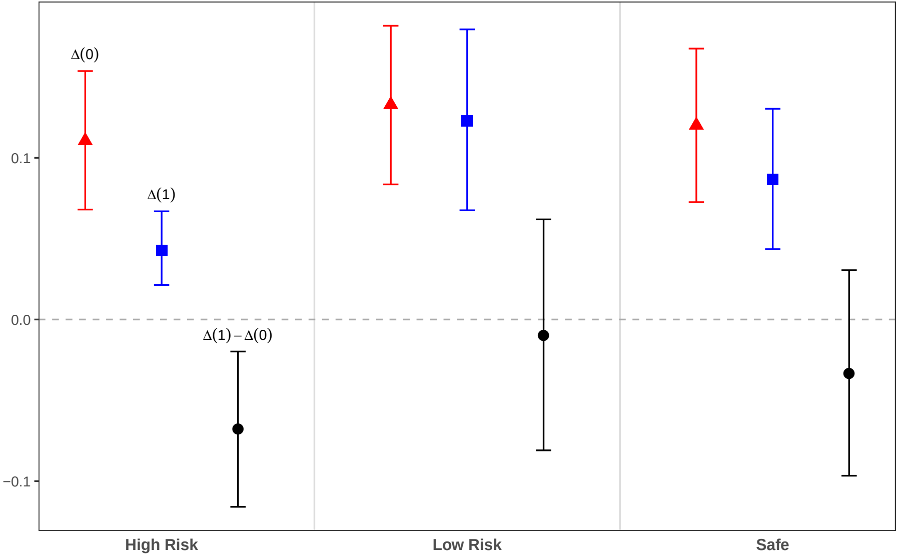
for $r = 0, \ldots, k-1$, and

$$\begin{aligned}\Delta_k(z) \;=\; \max_{s,s'} \big| & \Pr\{D_i(z) = k \mid S_i = s, R_i = r\} \\ & - \;\Pr\{D_i(z) = k \mid S_i = s', R_i = r\}\big|\end{aligned}$$

- Does the provision of PRAI improve the fairness of judges' decision?

$$\Delta_r(1) - \Delta_r(0), \quad \Delta_k(1) - \Delta_k(0)$$

Estimated Measure of Fairness



Optimal Decision Rule

- Can experimental data help judges achieve their goal?
- Goal: prevent as many NCA as possible with the least amount of bail
- Judge's decision rule:

$$\delta : \mathcal{X} \rightarrow \{0, 1, \ldots, k\}$$

  where $\mathcal{X}$ is the support of X_i, which may include PRAI
- $0 - 1$ utility:

$$1\{\delta(\mathsf{X}_i) = R_i\}$$

- Maximize the expected utility

$$\delta^* \quad = \quad \underset{\delta}{\mathrm{argmax}}\, \mathbb{E}[1\{\delta(\mathsf{X}_i) = R_i\}] \; = \; \underset{r \in \{0,1,\ldots,k\}}{\mathrm{argmax}} \; e_r(\mathsf{x})$$

- Optimal decision is not necessarily fair

Optimal PRAI Provision Rule

- Judges may not follow the above recommendation
- Policymakers can decide when to provide PRAI to judges
- The experiment cannot tell what is the optimal PRAI
  $\rightsquigarrow$ this requires the randomization of PRAI itself!
- PRAI provision rule:
  $$\xi : \mathcal{X} \to \{0, 1\}$$
- Judge's decision can randomly vary across cases with the same covariate values: $\Pr(\delta_{iz}(\mathrm{x}) = d) = \Pr(\delta_{i'z}(\mathrm{x}) = d)$
- $0 - 1$ utility:
  $$1\{\delta_{i,\xi(\mathrm{X}_i)} = R_i\}$$
- Maximize the expected utility
  $$\begin{aligned} \xi(\mathrm{x}) &= \operatorname{argmax} \mathbb{E}[1\{\delta_{i,\xi(\mathrm{X}_i)} = R_i\}] \\ &= \operatorname*{argmax}_z \sum_{r=0}^{k} e_r(\mathrm{x}) \cdot \Pr(D_i = r \mid Z_i = z, \mathrm{X}_i). \end{aligned}$$

Concluding Remarks

- We offer a set of statistical methods for experimentally evaluating computer-assisted human decision making
  1. causal quantities of interest based on principal stratification
  2. partial identification with a minimal set of assumptions
  3. point identification under unconfoundedness
  4. estimation strategies based on principal score weighting
  5. sensitivity analysis
  6. optimal decision rule
  7. optimal machine-recommendation provision rule
  8. fairness of decision based on principal stratification
- Development of an open-source software package
- Application to pretrial risk assessment instrument
  - first field experiment since the 1981–82 Philadelphia experiment
  - empirical analysis is currently underway