Experimental Evaluation of Algorithm-Assisted Human Decision Making: Application to Pretrial Public Safety Assessment

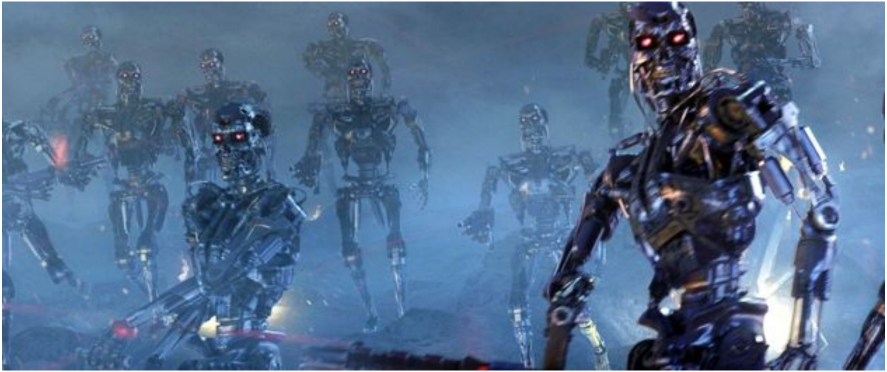
Kosuke Imai

Harvard University

Microsoft Research, New England
October 29, 2020

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

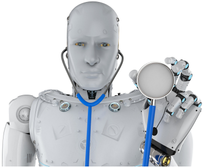
Rise of the Machines



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

Algorithm-Assisted Human Decision Making

- But, humans still make many consequential decisions
- We have not yet outsourced these decisions to machines



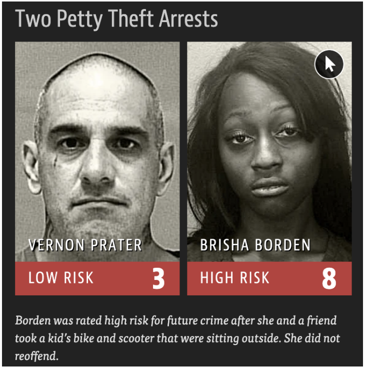
  - this is true even when human decisions can be suboptimal
  - we may want to hold *someone*, rather than *something*, accountable

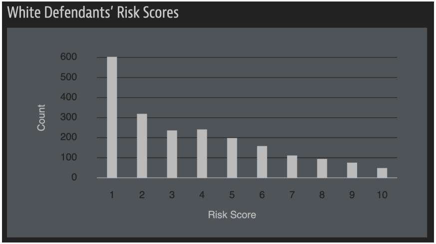
- Most prevalent system is algorithm-assisted human decision making
  - humans make decisions with the aid of algorithmic recommendations
  - routine decisions made by individuals in daily lives
  - consequential decisions made by judges, doctors, etc.

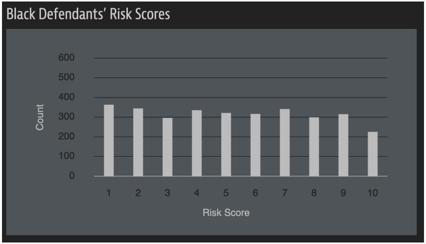
Questions and Contributions

- How do algorithmic recommendations influence human decisions?
  - Do they help human decision-makers achieve their goal?
  - Do they help humans improve the fairness of their decisions?
- Many have studied the accuracy and fairness of algorithms
  - Few have researched their impacts on human decisions
  - Little is known about how algorithmic bias interacts with human bias
- Our contributions:
  1. experimental evaluation of algorithm-assisted human decision making
  2. principal fairness: new fairness notion based on causality
  3. real-world field experiment evaluating pretrial public safety assessment

Controversy over the COMPAS Score (Propublica)







Pretrial Public Safety Assessment (PSA)

- Algorithmic 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 while preserving public safety
- Judges are required to consider three risk factors along with others
  1. arrestee may fail to appear in court (FTA)
  2. arrestee may engage in new criminal activity (NCA)
  3. arrestee may engage in new violent criminal activity (NVCA)
- PSA as an algorithmic recommendation to judges
  - classifying arrestees according to FTA and NCA/NVCA risks
  - derived from an application of a machine learning algorithm to a training data set based on past observations
  - different from COMPAS score

A Field Experiment for Evaluating the PSA

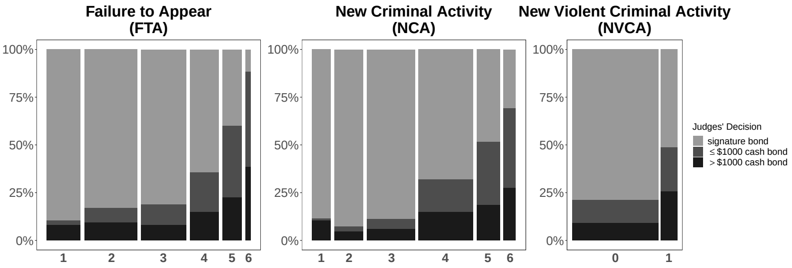
- Dane County, Wisconsin
- PSA = weighted indices of ten factors
  1. two separate ordinal six-point risk scores for FTA and NCA
  2. one binary risk score for new violent criminal activity (NVCA)
  3. age as the single demographic factor: no gender or race
  4. nine factors drawn from criminal history (prior convictions and FTA)
- Judges may 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)
- Field experiment
  - clerk assigns case numbers sequentially as cases enter the system
  - PSA is calculated for each case using a computer system
  - if the first digit of case number is even, PSA is given to the judge
  - mid-2017 – 2019 (randomization), 2-year follow-up for half sample

PSA Provision, Demographics, and Outcomes

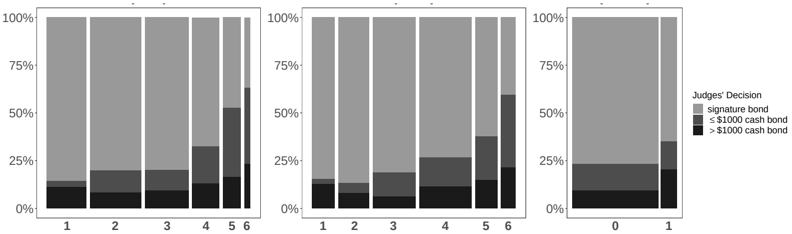
	no PSA			PSA			
	Signature	Cash bond		Signature	Cash bond		
	bond	small	large	bond	small	large	Total (%)
Non-white female	64	11	6	67	6	0	154 (8)
White female	91	17	7	104	17	10	246 (13)
Non-white male	261	56	49	258	53	57	734 (39)
White male	289	48	44	276	54	46	757 (40)
FTA committed	218	42	16	221	45	16	558 (29)
not committed	487	90	90	484	85	97	1333 (71)
NCA committed	211	39	14	202	40	17	523 (28)
not committed	494	93	92	503	90	96	1368 (72)
NVCA committed	36	10	3	44	10	6	109 (6)
not committed	669	122	103	661	120	107	1782 (94)
Total (%)	705	132	106	705	130	113	1891
	(37)	(7)	(6)	(37)	(7)	(6)	(100)

Judge's Decision Is Positively Correlated with PSA

(a) Treatment Group

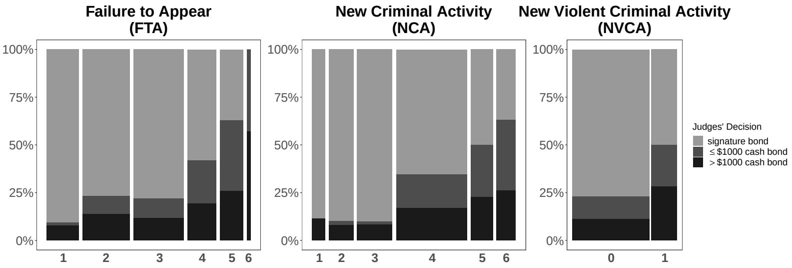


(b) Control Group

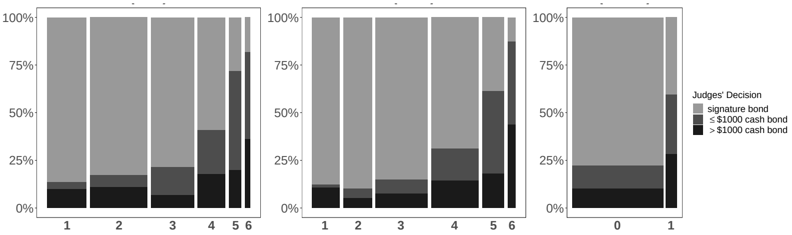


Racial Differences between Non-white and White Males

(a) Non-White Males

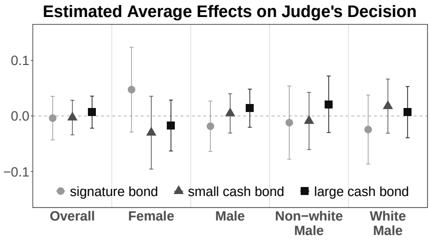


(b) White Males

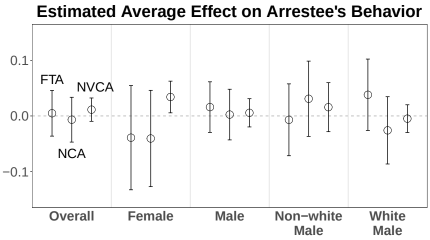


Intention-to-Treat Analysis of PSA Provision

(a) Estimated effects on judge's decisions



(b) Estimated effects on outcomes



- Difference-in-means estimator
- Insignificant effects on judge's decisions
- Possible effect on NVCA outcome for females
- Need to explore causal heterogeneity based on risk-levels

The Setup of the Proposed Methodology

- Notation:
  - $i = 1, 2, \ldots, n$: cases
  - Z_i: whether PSA is presented to the judge ($Z_i = 1$) or not ($Z_i = 0$)
  - D_i: judge's binary decision to detain ($D_i = 1$) or release ($D_i = 0$)
  - 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: we 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: $Y_i(1) \leq Y_i(0)$ for all i

Causal Quantities of Interest

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

- Average principal causal effects of PSA on judge's decisions:

$$\begin{aligned} \text{APCEp} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(0) = 1, Y_i(1) = 0\}, \\ \text{APCEr} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(0) = 1, Y_i(1) = 1\}, \\ \text{APCEs} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(0) = 0, Y_i(1) = 0\}. \end{aligned}$$

- If PSA is helpful, we should have APCEp > 0 and APCEs < 0
- The desirable sign of APCEr depends on various factors

Partial Identification Results

- The assumptions of randomization, exclusion restriction, and monotonicity imply,

$$\begin{aligned}
\text{APCEp} &= \frac{\Pr(Y_i = 1 \mid Z_i = 0) - \Pr(Y_i = 1 \mid Z_i = 1)}{\Pr\{Y_i(0) = 1\} - \Pr\{Y_i(1) = 1\}}, \\
\text{APCEr} &= \frac{\Pr(D_i = 1, Y_i = 1 \mid Z_i = 1) - \Pr(D_i = 1, Y_i = 1 \mid Z_i = 0)}{\Pr\{Y_i(1) = 1\}}, \\
\text{APCEs} &= \frac{\Pr(D_i = 0, Y_i = 0 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 1)}{1 - \Pr\{Y_i(0) = 1\}}.
\end{aligned}$$

- The signs of APCE are identifiable
- The bounds on APCE can be obtained

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

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

$$\begin{aligned}\text{APCEp} &= \mathbb{E}\{w_P(\mathrm{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_P(\mathrm{X}_i)D_i \mid Z_i = 0\},\\ \text{APCEr} &= \mathbb{E}\{w_R(\mathrm{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_R(\mathrm{X}_i)D_i \mid Z_i = 0\},\\ \text{APCEs} &= \mathbb{E}\{w_S(\mathrm{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_S(\mathrm{X}_i)D_i \mid Z_i = 0\},\end{aligned}$$

where

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

and

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

Extension to Ordinal Decision

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

$$R_i \ = \ \begin{cases} \min\{d : Y_i(d) = 0\} & \text{if } Y_i(k) = 0 \\ k+1 & \text{if } Y_i(k) = 1 \end{cases}$$

  - Least amount of bail that keeps an arrestee from committing NCA
  - Example with $k = 2$: risky cases ($R_i = 3$), preventable cases ($R_i = 2$ and $R_i = 1$), safe cases ($R_i = 0$)
- Causal quantities of interest: reduction in the proportion of NCA attributable to the PSA within each principal strata $r = 1, \ldots, k$

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

- Nonparametric identification under unconfoundedness

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

$$\begin{aligned} D_i^*(z) &= \beta_Z z + \mathrm{X}_i^\top \beta_X + z\mathrm{X}_i^\top \beta_{zx} + \epsilon_{i1}, \\ R_i^* &= \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2}, \end{aligned}$$

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

- Under unconfoundedness, we have

$$Y_i^* = \sum_{r=0}^{k} \delta_r 1(D_i = r) - \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2},$$

  where $Y_i = 1$ if $Y_i^* > 0$ and $Y_i = 0$ if $Y_i^* \leq 0$.

- Bayesian computation via MCMC

Principal Fairness (Imai and Jiang, 2020)

- Literature focuses on the fairness of algorithmic recommendations
- We focus on the fairness of human decision

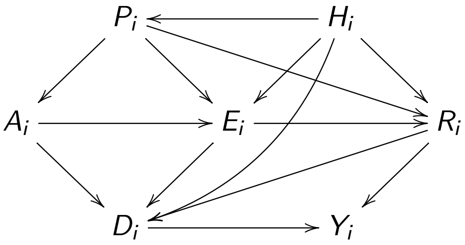
- Principal fairness: decision should not (statistically) depend on a protected attribute A_i (e.g., race and gender) within a principal strata

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

- Existing statistical fairness definitions do not take into account how a decision affects individuals
  1. Overall parity: $D_i \perp\!\!\!\perp A_i$
  2. Calibration: $Y_i \perp\!\!\!\perp A_i \mid D_i$
  3. Accuracy: $D_i \perp\!\!\!\perp A_i \mid Y_i$
- These three criteria may not hold simultaneously

Relationships with the Existing Statistical Fairness Criteria

- All groups are created equal: There exist a set of covariates W_i such that the principal strata are conditionally independent of the protected attribute given W_i, i.e., $R_i \perp\!\!\!\perp A_i \mid W_i$.



- H_i: historical processes
- P_i: parents' characteristics
- E_i: socio-economic factors

- Under this assumption, principal fairness implies all the other criteria

Measuring and Estimating the Degree of Fairness

- How fair are the judges' decisions?

$$\Delta_r(z) = \max_{a,a',d} \Big| \Pr\{D_i(z) \geq d \mid A_i = a, R_i = r\} - \Pr\{D_i(z) \geq d \mid A_i = a' R_i = r\} \Big|$$

for $1 \leq d \leq k$ and $0 \leq r \leq k+1$

- Does the provision of PSA improve the fairness of the judge's decision?

$$\Delta_r(1) - \Delta_r(0)$$

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:

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

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

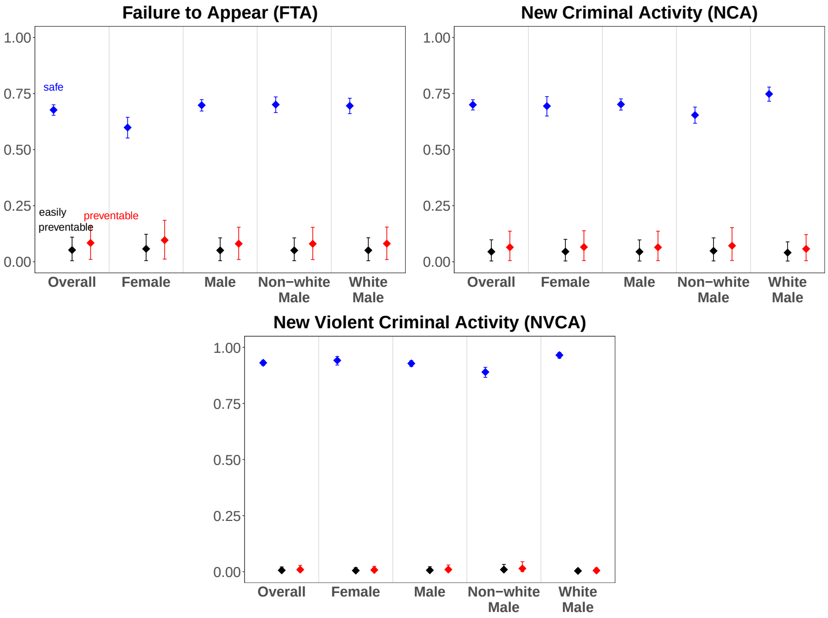
$$1\{\delta(\mathsf{X}_i) = R_i\}$$

- Maximize the expected utility

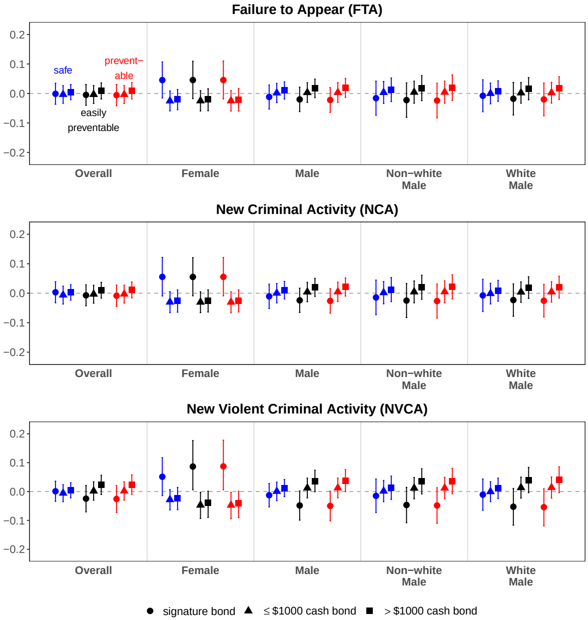
$$\delta^* \;=\; \operatorname*{argmax}_{\delta} \mathbb{E}[1\{\delta(\mathsf{X}_i) = R_i\}] \;=\; \operatorname*{argmax}_{r \in \{0,1,\ldots,k\}} e_r(\mathsf{x})$$

- Optimal decision is not necessarily fair

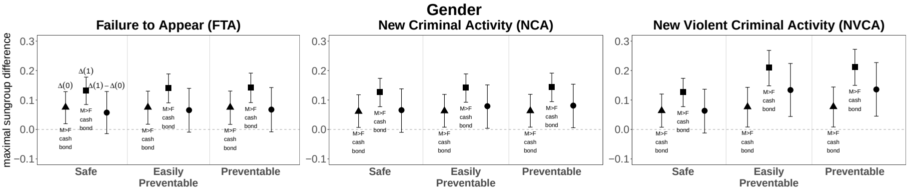
Estimated Proportion of Principal Strata

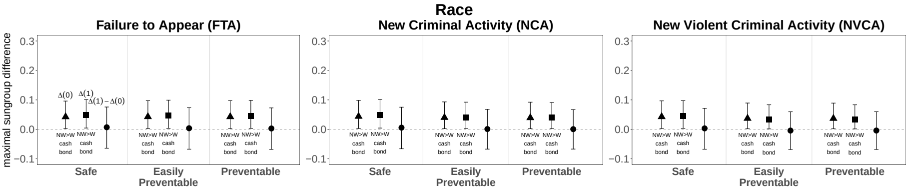


Estimated Average Principal Causal Effects



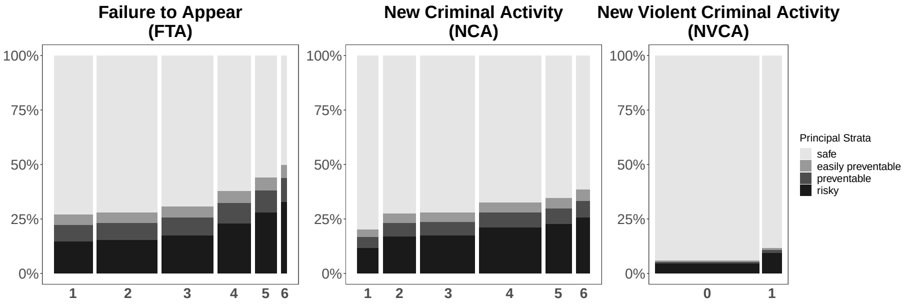
Principal Fairness





Optimal Decision and PSA

- Recall that the highest principal score gives the optimal decision
- For all observed covariates in the data, the optimal decision is to impose a signature bond
- Comparison between PSA and principal score



Concluding Remarks

- We offer a set of statistical methods for experimentally evaluating algorithm-assisted human decision making
- Field experiment for assessing the pretrial public safety assessment
  - most existing research uses observational data or hypothetical survey experiment
  - first field experiment since the small 1981–82 Philadelphia experiment about a new bond guideline
  - more ongoing experiments in this and several other counties
- Ongoing research
  - extension to multi-dimensional decision
  - role of incarceration
  - optimal PSA provision vs. optimal PSA
  - effects of PSA on judges and arrestees over time