

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

Kosuke Imai    Zhichao Jiang  
D. James Greiner    Ryan Hallen    Sooahn Shin

Harvard University  
University of Massachusetts, Amherst

Discussion Paper Meeting, Research Section  
Royal Statistical Society  
February 8, 2022

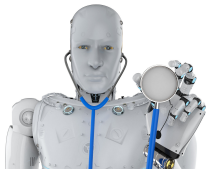
# Rise of the Machines



- 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



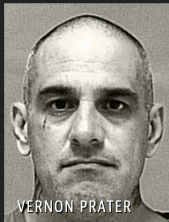
- this is true even when human decisions can be suboptimal
- we may want to hold *someone*, rather than *something*, accountable
- Most prevalent system is **algorithm-assisted human decision making**
  - humans make decisions with the aid of algorithmic recommendations
  - routine decisions made by 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 **methodology**: causal inference, fairness, and optimal decision
  - 3 **first ever field experiment** evaluating pretrial public safety assessment

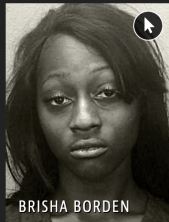
# Controversy over the COMPAS Score (Propublica)

## Two Petty Theft Arrests



LOW RISK

3

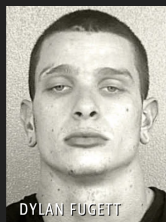


HIGH RISK

8

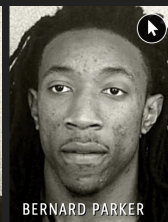
*Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.*

## Two Drug Possession Arrests



LOW RISK

3

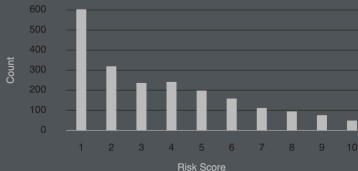


HIGH RISK

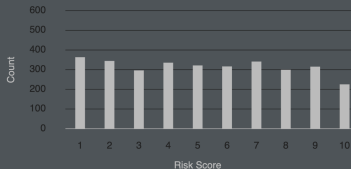
10

*Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.*

## White Defendants' Risk Scores



## Black Defendants' Risk Scores



# 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
  - age as the single demographic factor: no gender or race
  - nine factors drawn from criminal history (prior convictions and FTA)
- PSA scores and recommendation
  - 1 two separate ordinal six-point risk scores for FTA and NCA
  - 2 one binary risk score for new violent criminal activity (NVCA)
  - 3 aggregate recommendation: signature bond, small and large cash bond
- 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



**DANE COUNTY CLERK OF COURTS**  
**Public Safety Assessment – Report**

215 S Hamilton St #1000  
Madison, WI 53703  
Phone: (608) 266-4311

Name: [REDACTED]

Spillman Name Number: [REDACTED]

DOB: [REDACTED]

Gender: Male

Arrest Date: 03/25/2017

PSA Completion Date: 03/27/2017

**New Violent Criminal Activity Flag**

No

**New Criminal Activity Scale**

1	2	3	4	5	6
---	---	---	---	---	---

**Failure to Appear Scale**

1	2	3	4	5	6
---	---	---	---	---	---

**Charge(s):**

961.41(1)(D)(1) MFC DELIVER HEROIN <3 GMS F 3

**Risk Factors:**

**Responses:**

1. Age at Current Arrest	23 or Older
2. Current Violent Offense	No
a. Current Violent Offense & 20 Years Old or Younger	No
3. Pending Charge at the Time of the Offense	No
4. Prior Misdemeanor Conviction	Yes
5. Prior Felony Conviction	Yes
a. Prior Conviction	Yes
6. Prior Violent Conviction	2
7. Prior Failure to Appear Pretrial in Past 2 Years	0
8. Prior Failure to Appear Pretrial Older than 2 Years	Yes
9. Prior Sentence to Incarceration	Yes

**Recommendations:**

Release Recommendation - Signature bond

Conditions - Report to and comply with pretrial supervision

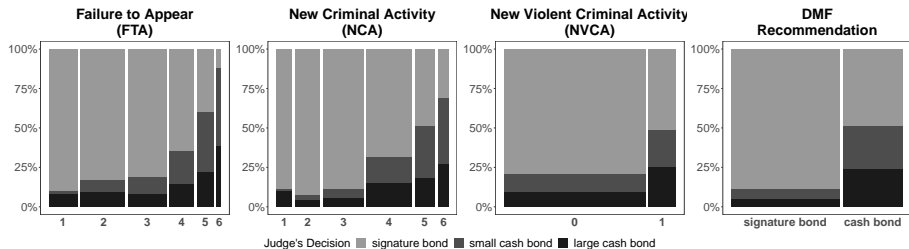


## PSA Provision, Demographics, and Outcomes

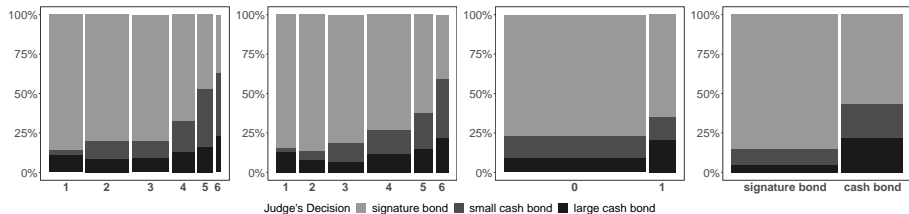
	no PSA			PSA			Total (%)
	Signature bond	Cash bond <i>small</i>	Cash bond <i>large</i>	Signature bond	Cash bond <i>small</i>	Cash bond <i>large</i>	
Non-white female	64	11	6	67	6	0	154 (8)
White female	91	17	7	104	17	10	246 (13)
Non-white male	261	56	49	258	53	57	734 (39)
White male	289	48	44	276	54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

# Judge's Decision Is Positively Correlated with PSA

## (a) Treatment Group

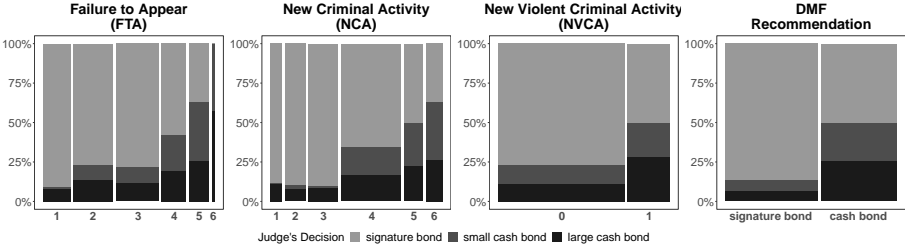


## (b) Control Group

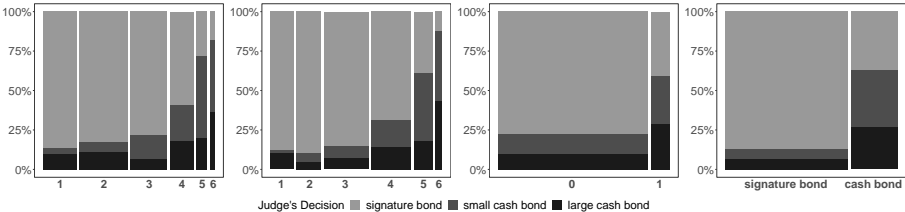


# Racial Differences between Non-white and White Males

(a) Non-White Males

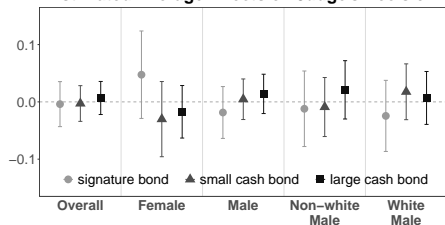


(b) White Males

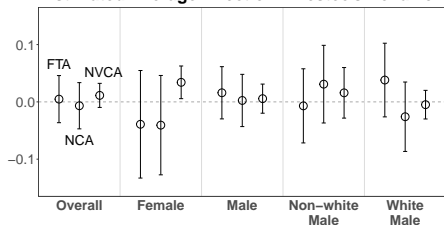


# Intention-to-Treat Analysis of PSA Provision

Estimated Average Effects on Judge's Decision



Estimated Average Effect on Arrestee's Behavior



- Difference-in-means estimator
- Insignificant effects on judge's decisions
- Possible effect on NVCA outcome for females

# How Does PSA Provision Influence Judges' Decisions?

- 1 Does PSA provision help prevent FTA, NCA, and NVCA?  
↪ ITT effects on the arrestees' behavior
- 2 Does PSA provision help judges make better decisions?
  - ITT effects on the judges' decision ↪ answers whether PSA provision makes decisions harsher or more lenient
  - a good decision: detain risky arrestees, release safe arrestees
  - need to explore causal heterogeneity based on **risk levels**

# The Setup of the Proposed Methodology (Binary Decision)

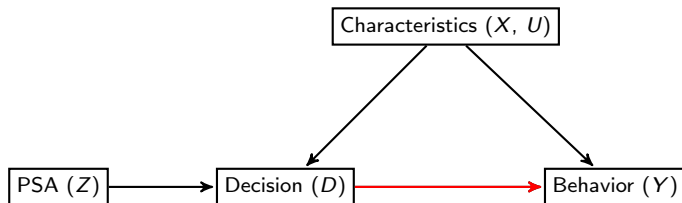
- Notation

- $Z_i$ : PSA provision indicator
- $D_i$ : detain ( $D_i = 1$ ) or release ( $D_i = 0$ )
- $Y_i$ : binary outcome (e.g., NCA)
- $X_i$ : observed covariates
- $U_i$ : unobserved covariates

- Potential outcomes

- $D_i(z)$ : potential value of the decision when  $Z_i = z$
- $Y_i(z, d)$ : potential outcome when  $Z_i = z$  and  $D_i = d$
- No interference across cases: first arrests only

# Assumptions



- **Randomized treatment assignment:**  $\{D_i(z), Y_i(z, d), X_i, U_i\} \perp\!\!\!\perp Z_i$
- **Exclusion restriction:**  $Y_i(z, d) = Y_i(d)$
- **Monotonicity:**  $Y_i(0) \geq Y_i(1)$

# Causal Quantities of Interest

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

- Average principal causal effects of PSA on judges' decisions:

$$APCE_p = \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 0, Y_i(0) = 1\},$$

$$APCE_r = \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 1, Y_i(0) = 1\},$$

$$APCE_s = \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 0, Y_i(0) = 0\}.$$

- If PSA is helpful, we should have  $APCE_p > 0$  and  $APCE_s < 0$ .
- The desirable sign of  $APCE_r$  depends on various factors.



## Partial Identification Results

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

$$\text{APCE}_p = \frac{\Pr(Y_i = 1 \mid Z_i = 0) - \Pr(Y_i = 1 \mid Z_i = 1)}{\Pr\{Y_i(0) = 1\} - \Pr\{Y_i(1) = 1\}}$$

$$\text{APCE}_r = \frac{\Pr(D_i = 1, Y_i = 1 \mid Z_i = 1) - \Pr(D_i = 1, Y_i = 1 \mid Z_i = 0)}{\Pr\{Y_i(1) = 1\}}$$

$$\text{APCE}_s = \frac{\Pr(D_i = 0, Y_i = 0 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 1)}{1 - \Pr\{Y_i(0) = 1\}}$$

- 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

- **Unconfoundedness:**  $Y_i(d) \perp\!\!\!\perp D_i \mid X_i, Z_i = z$
- Violation of unconfoundedness
  - unobserved covariates between decision and outcome
  - sensitivity analysis
- **Principal score**

$$e_P(x) = \Pr\{Y_i(1) = 1, Y_i(0) = 0 \mid X_i = x\}$$

- Identification formula

$$\text{APCE}_P = \mathbb{E} \left[ \underbrace{\frac{e_P(x)}{\mathbb{E}\{e_P(X_i)\}}}_{\text{weight}} D_i \mid Z_i = 1 \right] - \mathbb{E} \left[ \underbrace{\frac{e_P(x)}{\mathbb{E}\{e_P(X_i)\}}}_{\text{weight}} D_i \mid Z_i = 0 \right]$$

## Extension to Ordinal Decision

- Judges decisions are typically ordinal (e.g., bail amount)
  - $D_i = 0, 1, \dots, k$ : a bail of increasing amount
  - Monotonicity**:  $Y_i(d_1) \geq Y_i(d_2)$  for  $d_1 \leq d_2$
- 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  $k = 2$

principal strata	$(Y_i(0), Y_i(1), Y_i(2))$	$R_i$
risky cases	(1, 1, 1)	3
preventable cases	(1, 1, 0)	2
easily preventable cases	(1, 0, 0)	1
safe cases	(0, 0, 0)	0

## APCE for Ordinal Decision

- For people with  $R_i = r$ 
  - judges make decision  $D_i \geq r \rightsquigarrow$  not commit a crime
  - judges make decision  $D_i < r \rightsquigarrow$  commit a crime
- **Causal quantities of interest** : reduction in the proportion of NCA attributable to PSA provision

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

- Nonparametric identification under unconfoundedness

## Parametric Model and Sensitivity Analysis

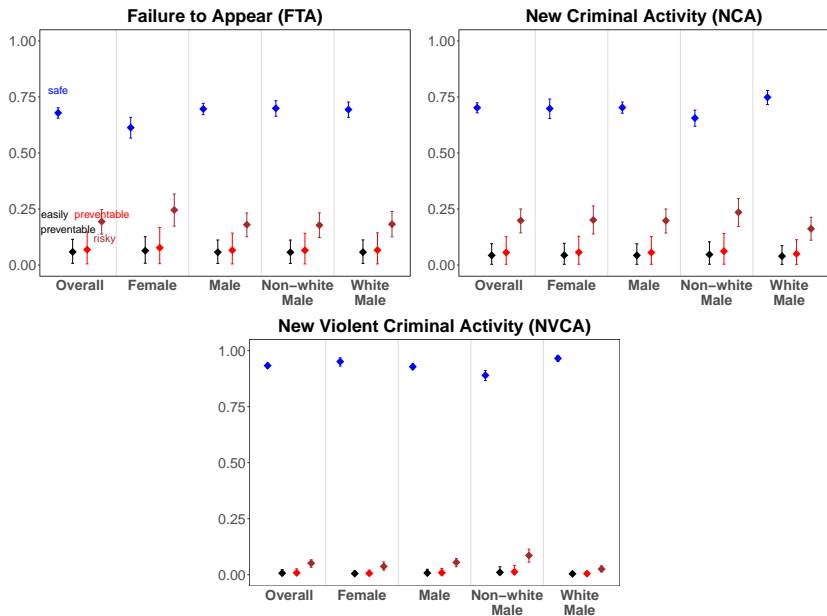
- Judges may use additional information when making decisions
- Sensitivity analysis: How robust are one's empirical results to the potential violation of the key assumption?
- Ordinal probit models for  $D_i(z)$  and  $R_i$  with latent variables

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

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

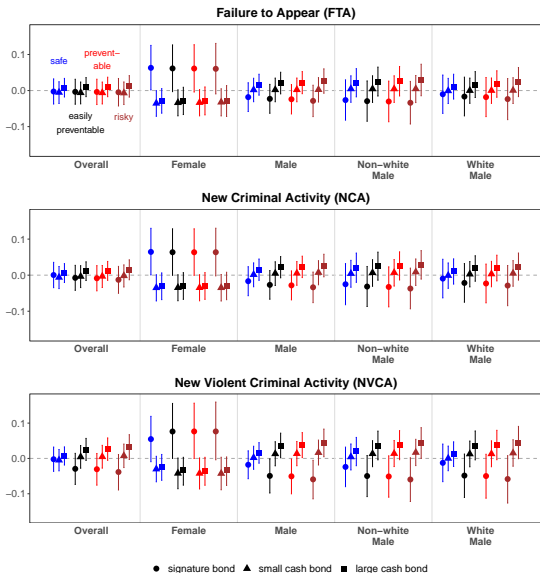
- $\rho = 0 \rightsquigarrow$  unconfoundedness

# Estimated Proportions of Principal Strata

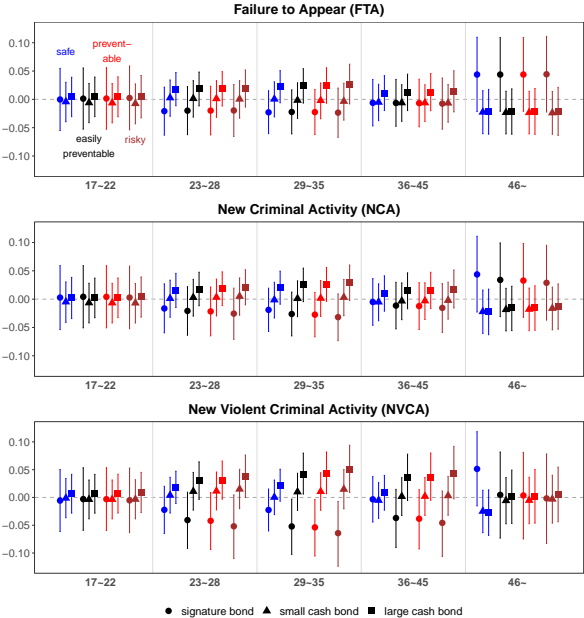


# Estimated Average Principal Causal Effects

$$\Pr\{D_i(1) = d \mid R_i = r\} - \Pr\{D_i(0) = d \mid R_i = r\}$$



# Estimates for Age Subgroups





# How Does PSA Provision Influence Judges' Decisions?

- 1 Does PSA provision help prevent FTA, NCA, and NVCA?  
↪ ITT effects on the arrestees' behavior
- 2 Does PSA provision help judges make better decisions?  
↪ Causal heterogeneity based on risk levels
- 3 Does PSA provision help judges improve the fairness of their decisions?

## Principal Fairness (Imai and Jiang, 2020)

- Literature focuses on the fairness of algorithmic recommendations
- We study the fairness of human decisions
- **Principal fairness:**  $D_i \perp\!\!\!\perp A_i \mid R_i = r$  for all  $r$ 
  - people with similar risk levels should be treated similarly
  - principal stratum fully characterizes the risk level
- 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:  $D_i \perp\!\!\!\perp A_i \mid Y_i$
- $R_i \perp\!\!\!\perp A_i \rightsquigarrow$  Principal fairness implies all statistical fairness criteria

## Measuring and Estimating the Degree of Fairness

- How fair are the judge's decisions?
- Between-group deviation in decision probability within each principal stratum

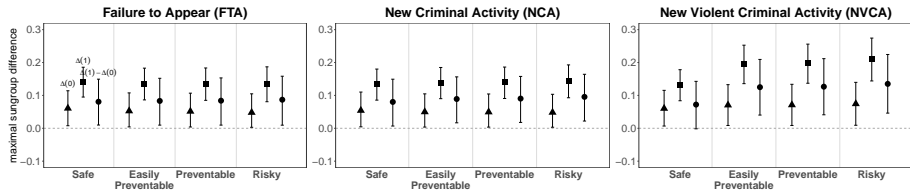
$$\Delta_r(z) = \max_{a, a', d} |\Pr\{D_i(z) \geq d \mid A_i = a, R_i = r\} \\ - \Pr\{D_i(z) \geq d \mid A_i = a', R_i = r\}|$$

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

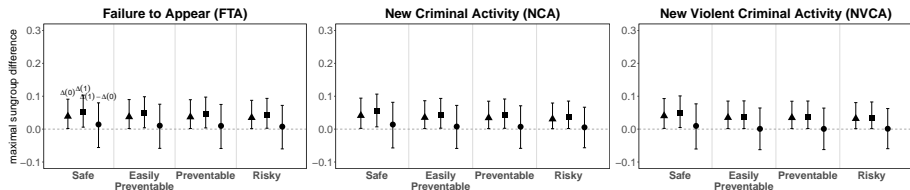
- Does the provision of PSA improve the fairness of the judge's decision?

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

# Gender and Racial Fairness



(a) Gender fairness



(b) Racial fairness

# How Does PSA Provision Influence Judges' Decisions?

- 1 Does PSA provision help prevent FTA, NCA, and NVCA?  
↪ ITT effects on the arrestees' behavior
- 2 Does PSA provision help judges make better decisions?  
↪ Causal heterogeneity based on risk levels
- 3 Does PSA provision help judges improve the fairness of their decisions?  
↪ Principal fairness
- 4 How should judges use  $X$  to make decisions?

# Optimal Decision Rule

- Goal: prevent as many NCA as possible with the least amount of bail
- Judge's decision rule:

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

where  $\mathcal{X}$  is the support of  $X_i$ , which may include PSA

- Utility:

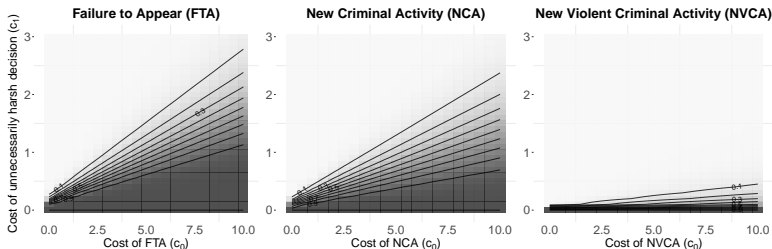
$$U_i(\delta) = \begin{cases} -c_0 & \delta(X_i) < R_i \quad (\text{too lenient}) \\ 1 & \delta(X_i) = R_i \\ 1 - c_1 & \delta(X_i) > R_i \quad (\text{unnecessarily harsh}) \end{cases}$$

where  $c_0, c_1 \geq 0$  are costs

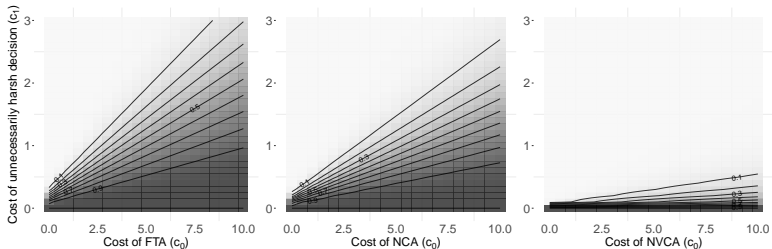
- Maximize the expected utility

$$\begin{aligned} \delta^* &= \operatorname{argmax}_{\delta} \mathbb{E}[U_i(\delta)] \\ &= \operatorname{argmax}_{r \in \{0, 1, \dots, k\}} \sum_{r \leq d} e_r(x) - c_0 \cdot \sum_{r > d} e_r(x) - c_1 \cdot \sum_{r < d} e_r(x). \end{aligned}$$

# Proportion of Cases for Which Cash Bond is Optimal



(a) The cases whose DMF recommendation is a signature bond



(b) The cases whose DMF recommendation is a cash bond

# Concluding Remarks

- We offer a set of statistical methods for experimentally evaluating algorithm-assisted human decision making
- Some potentially suggestive findings:
  - 1 little overall impacts on the judge's decisions
  - 2 more lenient decisions for females regardless of risk levels
  - 3 more stringent decisions for “risky” males
  - 4 widening gender bias, no effect on racial bias against non-whites
  - 5 signature bond is optimal unless the cost of new crime is high
- Ongoing research
  - more data, more experiments
  - improvement on the existing algorithm
  - dynamic interactions over time
  - data features: multi-dimensional decision, multi-site data, multiple cases