

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

Kosuke Imai Sooahn Shin

Harvard University

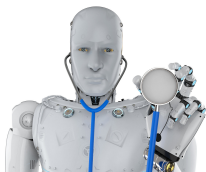
NYU Center for Experimental Social Science

March 5, 2021

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

Algorithm-Assisted Human Decision Making

- AI, machine learning, and statistics affect every aspect of our lives
- But, humans still make many consequential decisions
- We have not yet outsourced these decisions to machines



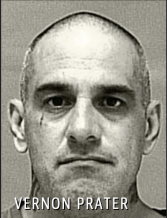
- this is true even when human decisions can be suboptimal
- we may want to hold *someone*, rather than *something*, accountable
- Most prevalent system is **algorithm-assisted human decision making**
 - humans make decisions with the aid of algorithmic recommendations
 - routine decisions made by individuals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

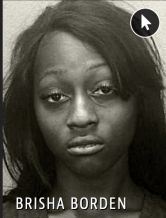
Controversy over the COMPAS Score (Propublica)

Two Petty Theft Arrests



VERNON PRATER

LOW RISK **3**




BRISHA BORDEN

HIGH RISK **8**

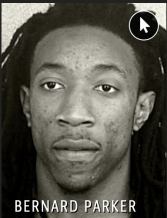
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Drug Possession Arrests



DYLAN FUGETT

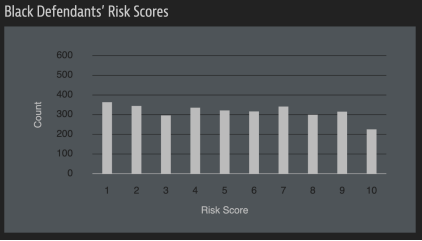
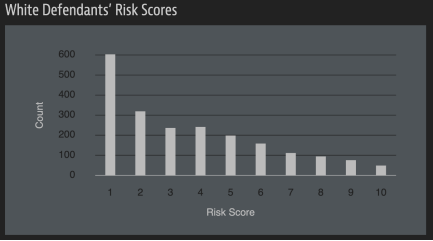
LOW RISK **3**



BERNARD PARKER

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.



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

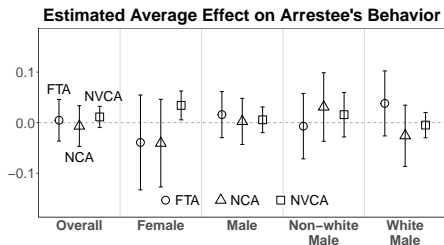
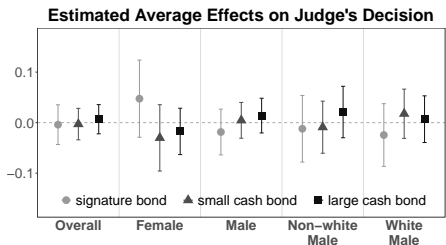
A Field Experiment for Evaluating the PSA

- Dane County, Wisconsin
- PSA = weighted indices of nine factors
 - age as the single demographic factor: no gender or race
 - eight 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
- 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			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)

Intention-to-Treat Analysis of PSA Provision



- 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 (Binary Decision)

- Notation:

- $i = 1, 2, \dots, n$: 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: $\{D_i(z), Y_i(z, d), X_i\} \perp\!\!\!\perp Z_i$
- 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 the judge's decisions:

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

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

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

- If PSA is helpful, we should have $\text{APCE}_p > 0$ and $\text{APCE}_s < 0$
- The desirable sign of APCE_r depends on various factors

Identification under Unconfoundedness

- **Unconfoundedness:**

$$Y_i(d) \perp\!\!\!\perp D_i \mid X_i, Z_i = z$$

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

- Violated if judges base their decision on additional information they have about arrestees \rightsquigarrow sensitivity analysis
- **Principal score:** population proportion of each principal stratum (Ding and Lu 2017)

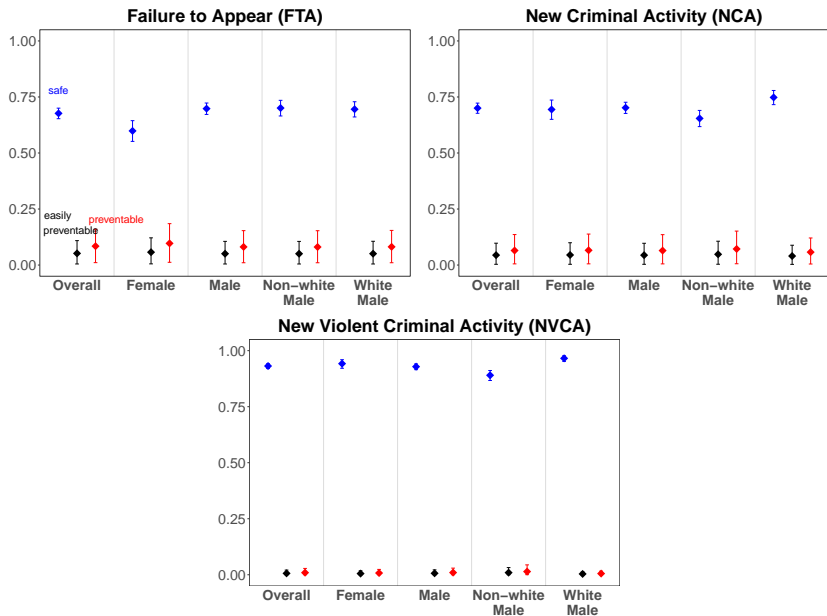
$$e_P(x) = \Pr\{Y_i(1) = 1, Y_i(0) = 0 \mid X_i = x\}$$

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

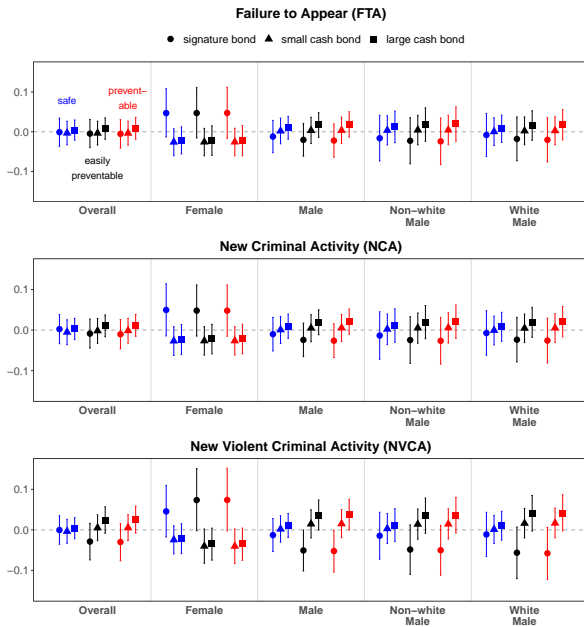
$$e_S(x) = \Pr\{Y_i(1) = 0, Y_i(0) = 0 \mid X_i = x\}$$

- Identification based on weighting with principal score
- Extension to ordinal decision is possible

Estimated Proportion of Principal Strata



Estimated Average Principal Causal Effects



Principal Fairness (Imai and Jiang, 2020)

- Literature focuses on the fairness of algorithmic recommendations
- We study the fairness of decisions by humans, algorithms, or humans with algorithmic recommendations
- **Principal fairness:** decision should not depend on a protected attribute A_i (e.g., race and gender) within a principal stratum

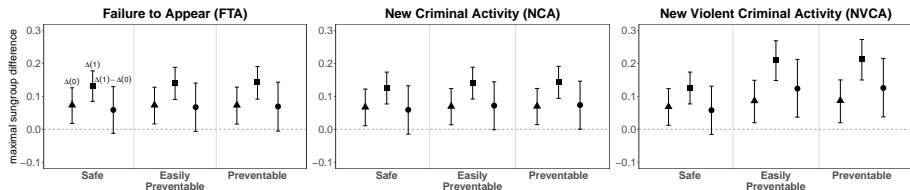
$$D_i \perp\!\!\!\perp A_i \mid R_i = r \quad \text{for all } r$$

- 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
- Under the “all groups are created equal” assumption, principal fairness (conditionally) implies all three fairness

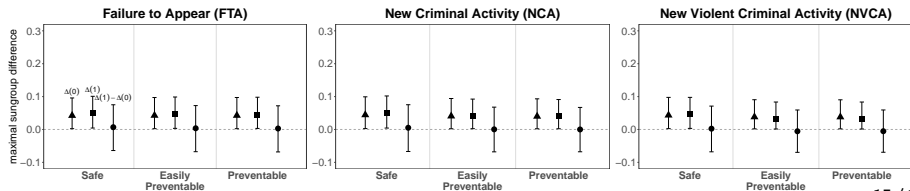
Gender and Racial Fairness

- Probability of imposing a signature vs. cash (small and large) bond

(a) Gender fairness



(b) Racial fairness



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
 - ⑥ judge's decisions may be too severe, PSA recommendation is more so
- Ongoing research
 - more data, more experiments
 - extension to multi-dimensional decision
 - role of incarceration
 - optimal PSA
 - effects of PSA on judges and arrestees over time