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

Kosuke Imai

Harvard University

The Japanese Society for Quantitative Political Science
January 9, 2023

Joint work with Zhichao Jiang, D. James Greiner,
Ryan Hallen, and Sooahn Shin

Algorithm-Assisted Human Decision Making

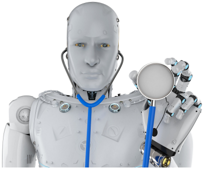
- Algorithms are everywhere in today's society

Algorithm-Assisted Human Decision Making

- Algorithms are everywhere in today's society
- But, humans still make many consequential decisions

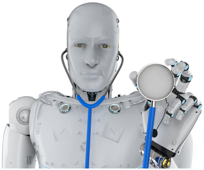
Algorithm-Assisted Human Decision Making

- Algorithms are everywhere in today's society
- But, humans still make many consequential decisions
- We have not yet outsourced these decisions to machines



Algorithm-Assisted Human Decision Making

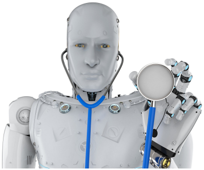
- Algorithms are everywhere in today's society
- But, humans still make many consequential decisions
- We have not yet outsourced these decisions to machines



  - this is true even when human decisions can be suboptimal

Algorithm-Assisted Human Decision Making

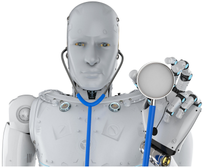
- Algorithms are everywhere in today's society
- But, humans still make many consequential decisions
- We have not yet outsourced these decisions to machines



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

Algorithm-Assisted Human Decision Making

- Algorithms are everywhere in today's society
- But, humans still make many consequential decisions
- We have not yet outsourced these decisions to machines

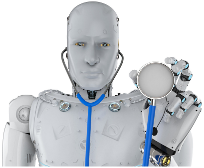


  - this is true even when human decisions can be suboptimal
  - we may want to hold *someone*, rather than *something*, accountable

- Most prevalent system is algorithm-assisted human decision making

Algorithm-Assisted Human Decision Making

- Algorithms are everywhere in today's society
- But, humans still make many consequential decisions
- We have not yet outsourced these decisions to machines

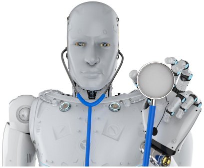


  - this is true even when human decisions can be suboptimal
  - we may want to hold *someone*, rather than *something*, accountable

- Most prevalent system is algorithm-assisted human decision making
  - humans make decisions with the aid of algorithmic recommendations

Algorithm-Assisted Human Decision Making

- Algorithms are everywhere in today's society
- But, humans still make many consequential decisions
- We have not yet outsourced these decisions to machines

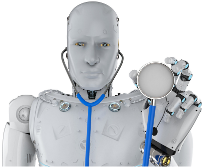



  - this is true even when human decisions can be suboptimal
  - we may want to hold *someone*, rather than *something*, accountable

- 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

Algorithm-Assisted Human Decision Making

- Algorithms are everywhere in today's society
- 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?

Questions and Contributions

- How do algorithmic recommendations influence human decisions?
  - Do they help human decision-makers achieve their goal?

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?

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

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

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

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:

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

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

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

Pretrial Public Safety Assessment (PSA)

- Algorithmic recommendations often used in US criminal justice system

Pretrial Public Safety Assessment (PSA)

- Algorithmic recommendations often used in US criminal justice system
- At the first appearance hearing, judges primarily make two decisions

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

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

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

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

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)

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)

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)

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

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

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

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

A Field Experiment for Evaluating the PSA

- Dane County, Wisconsin
- PSA $=$ weighted indices of ten factors

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

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)

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

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

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)

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

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

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

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)

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

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

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

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

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

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

DANE COUNTY CLERK OF COURTS
Public Safety Assessment – Report

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

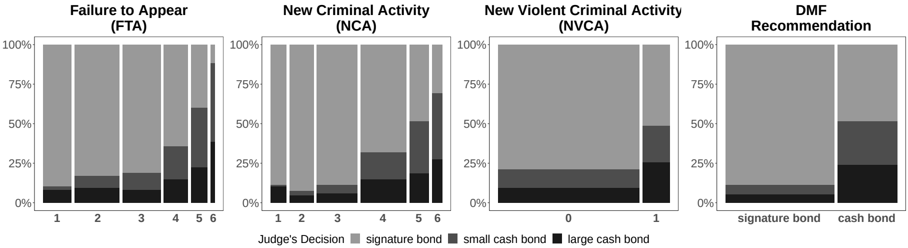
Conditions - Report to and comply with pretrial supervision

PSA Provision, Demographics, and Outcomes

	no PSA			PSA			
	Signature bond	Cash bond *small*	Cash bond *large*	Signature bond	Cash bond *small*	Cash bond *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 (37)	132 (7)	106 (6)	705 (37)	130 (7)	113 (6)	1891 (100)

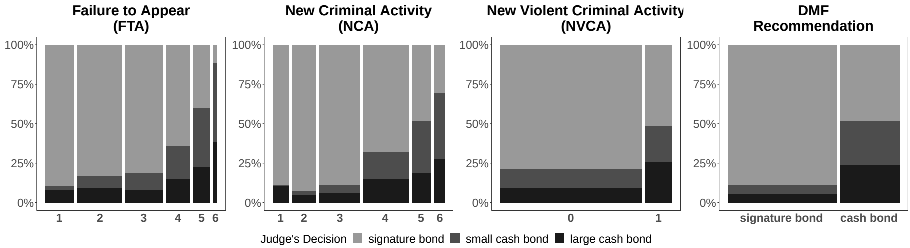
Judge's Decision Is Positively Correlated with PSA

(a) Treatment Group

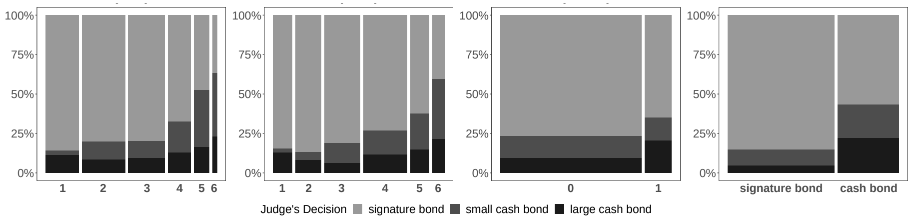


Judge's Decision Is Positively Correlated with PSA

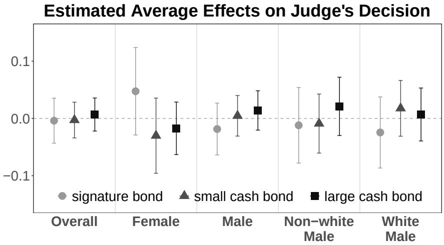
(a) Treatment Group



(b) Control Group

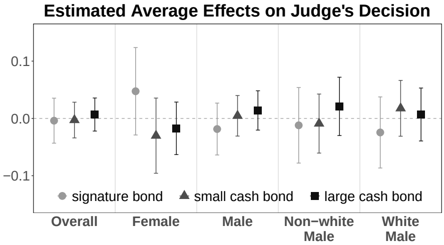


Intention-to-Treat Analysis of PSA Provision



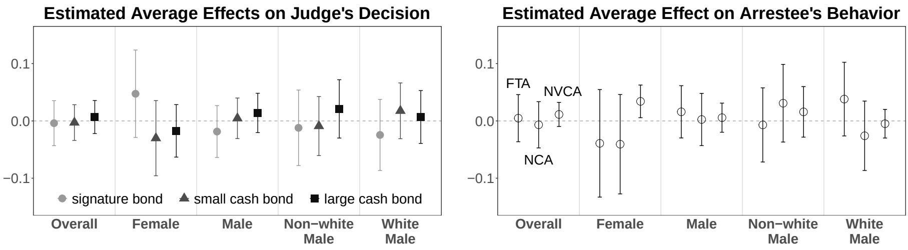
- Difference-in-means estimator

Intention-to-Treat Analysis of PSA Provision



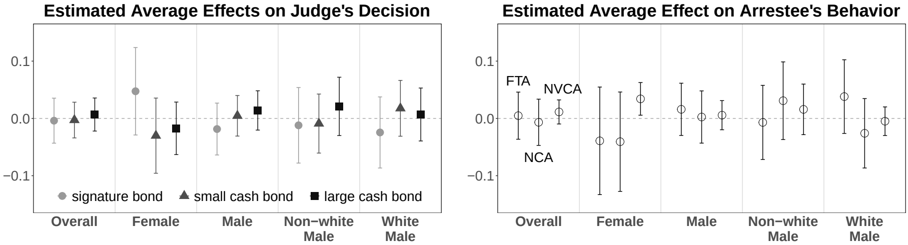
- Difference-in-means estimator
- Insignificant effects on judge's decisions

Intention-to-Treat Analysis of PSA Provision



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

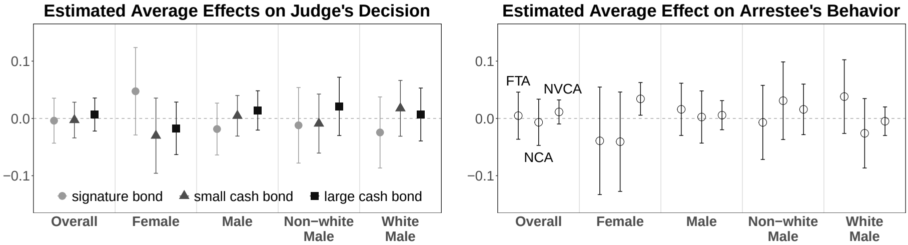
Intention-to-Treat Analysis of PSA Provision



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

- Does PSA provision help judges make "better" decisions?

Intention-to-Treat Analysis of PSA Provision



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

- Does PSA provision help judges make "better" decisions?
- Need to explore causal heterogeneity based on risk-levels

The Setup of the Proposed Methodology (Binary Decision)

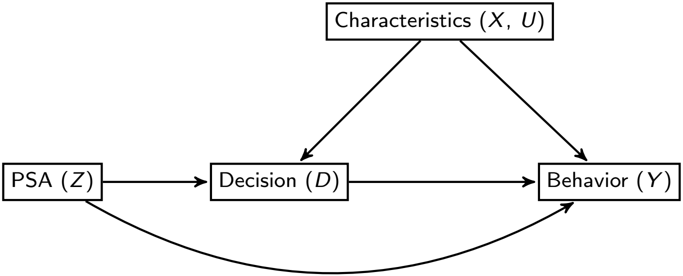
- Notation
  - Z_i: PSA provision indicator
  - D_i: detain ($D_i = 1$) or release ($D_i = 0$)
  - Y_i: binary outcome (e.g., NCA)
  - X_i: observed covariates
  - U_i: unobserved covariates

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

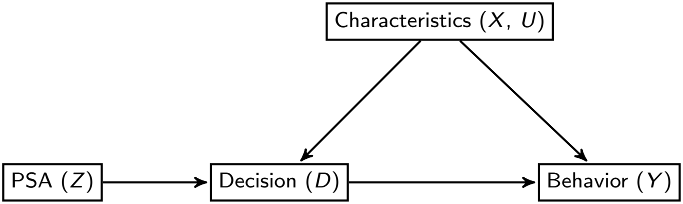
- Potential outcomes
  - $D_i(z)$: potential value of the decision when $Z_i = z$
  - $Y_i(z, d)$: potential outcome when $Z_i = z$ and $D_i = d$
  - No interference across cases: first arrests only

Assumptions



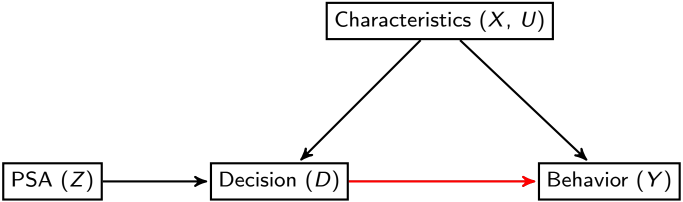
- Randomized treatment assignment: $\{D_i(z), Y_i(z,d), X_i, U_i\} \perp\!\!\!\perp Z_i$

Assumptions



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

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
  - $\cancel{(Y_i(1), Y_i(0)) = (1, 0)}$: eliminated by monotonicity

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
  - $\cancel{(Y_i(1), Y_i(0)) = (1, 0)}$: eliminated by monotonicity

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

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

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
  - $\cancel{(Y_i(1), Y_i(0)) = (1, 0)}$: eliminated by monotonicity

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

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

- 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

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

- Unconfoundedness: $Y_i(d) \perp\!\!\!\perp D_i \mid \mathrm{X}_i, Z_i = z$
- Violation of unconfoundedness
  - unobserved covariates between decision and outcome
  - sensitivity analysis

Point Identification

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

$$e_P(\mathrm{x}) \quad = \quad \Pr\{Y_i(1) = 1, Y_i(0) = 0 \mid \mathrm{X}_i = \mathrm{x}\}$$

Point Identification

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

$$e_P(\mathrm{x}) \quad = \quad \Pr\{Y_i(1) = 1, Y_i(0) = 0 \mid \mathrm{X}_i = \mathrm{x}\}$$

- Identification formula

$$\text{APCEp} \quad = \quad \mathbb{E}\left[\underbrace{\frac{e_P(\mathrm{x})}{\mathbb{E}\{e_P(\mathrm{X}_i)\}}}_{\text{weight}} D_i \mid Z_i = 1\right] - \mathbb{E}\left[\underbrace{\frac{e_P(\mathrm{x})}{\mathbb{E}\{e_P(\mathrm{X}_i)\}}}_{\text{weight}} D_i \mid Z_i = 0\right]$$

Extension to Ordinal Decision

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

Extension to Ordinal Decision

- Judges decisions are typically ordinal (e.g., bail amount)
  - $D_i = 0, 1, \ldots, k$: 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

Extension to Ordinal Decision

- Judges decisions are typically ordinal (e.g., bail amount)
  - $D_i = 0, 1, \ldots, k$: 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

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

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

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

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?

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

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

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

Parametric Model and Sensitivity Analysis

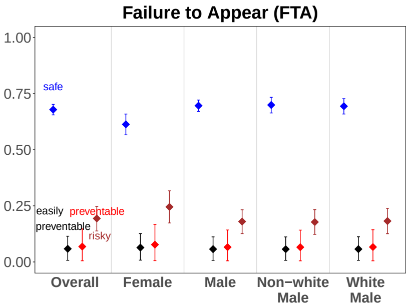
- Judges may use additional information when making decisions
- Sensitivity analysis: How robust are one's empirical results to the potential violation of the key assumption?
- Ordinal probit models for $D_i(z)$ and R_i with latent variables

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

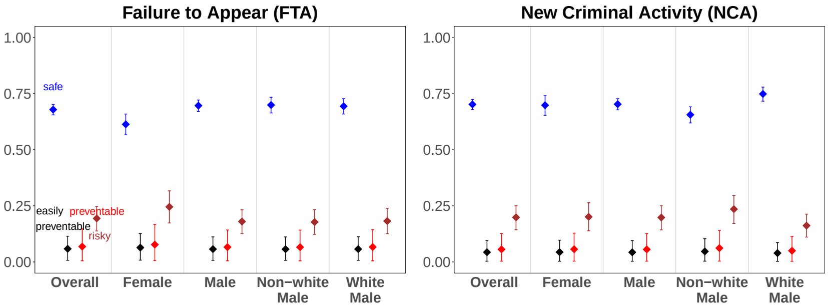
where $\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$.

- $\rho = 0 \rightsquigarrow$ unconfoundedness

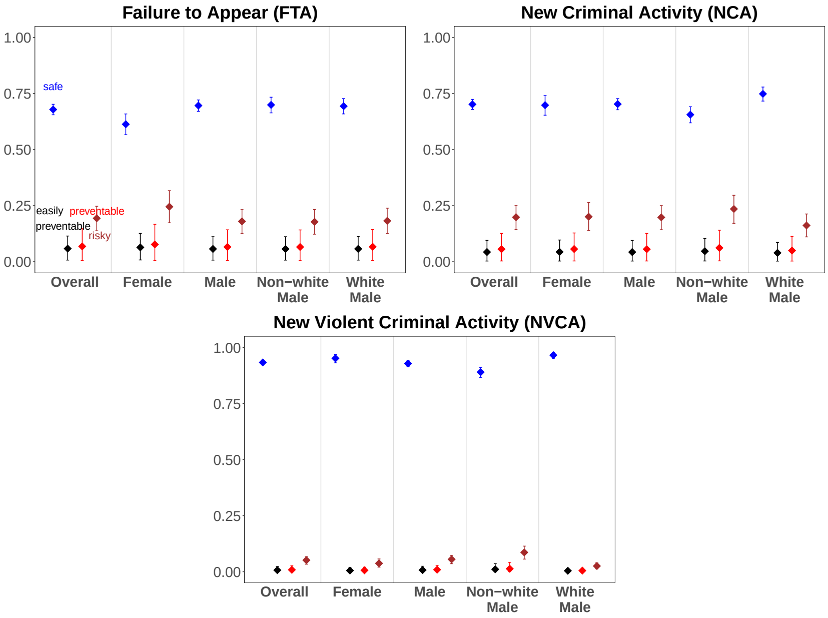
Estimated Proportions of Principal Strata



Estimated Proportions of Principal Strata

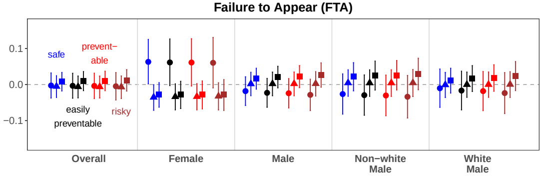


Estimated Proportions of Principal Strata



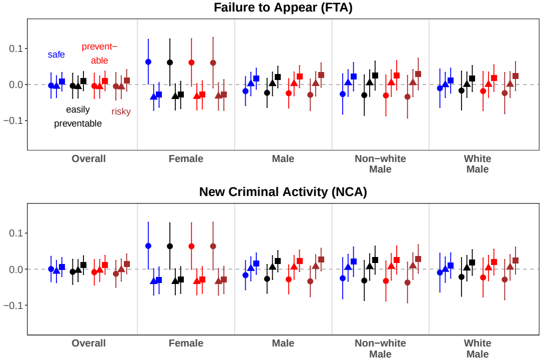
Estimated Average Principal Causal Effects

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



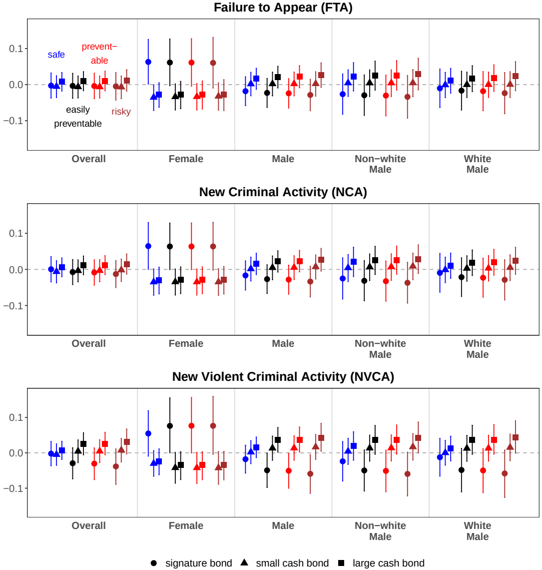
Estimated Average Principal Causal Effects

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



Estimated Average Principal Causal Effects

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



Does PSA Provision Improve Fairness of Judge's Decision?

- Literature focuses on the fairness of algorithmic recommendations

Does PSA Provision Improve Fairness of Judge's Decision?

- Literature focuses on the fairness of algorithmic recommendations
- We study the fairness of human decisions with algorithmic recommendation

Does PSA Provision Improve Fairness of Judge's Decision?

- Literature focuses on the fairness of algorithmic recommendations
- We study the fairness of human decisions with algorithmic recommendation
- Principal fairness (Imai and Jiang, in-press) :

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

  - people with similar risk levels should be treated similarly
  - principal stratum fully characterizes the risk level

Does PSA Provision Improve Fairness of Judge's Decision?

- Literature focuses on the fairness of algorithmic recommendations
- We study the fairness of human decisions with algorithmic recommendation
- Principal fairness (Imai and Jiang, in-press) :

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

  - 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: $D_i \perp\!\!\!\perp A_i \mid Y_i$

Does PSA Provision Improve Fairness of Judge's Decision?

- Literature focuses on the fairness of algorithmic recommendations
- We study the fairness of human decisions with algorithmic recommendation
- Principal fairness (Imai and Jiang, in-press) :

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

  - 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?

Measuring and Estimating the Degree of Fairness

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

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

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

Measuring and Estimating the Degree of Fairness

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

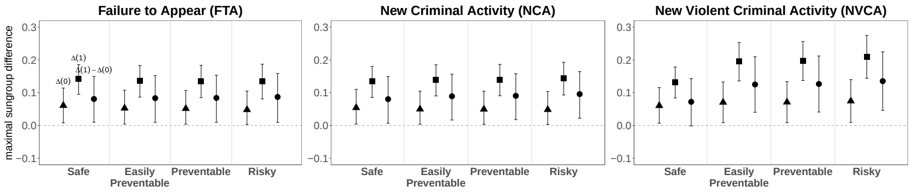
$$\Delta_r(z) = \max_{a,a',d} \Big| \Pr\{D_i(z) \geq d \mid A_i = a, R_i = r\} - \Pr\{D_i(z) \geq d \mid A_i = a', R_i = r\} \Big|$$

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?

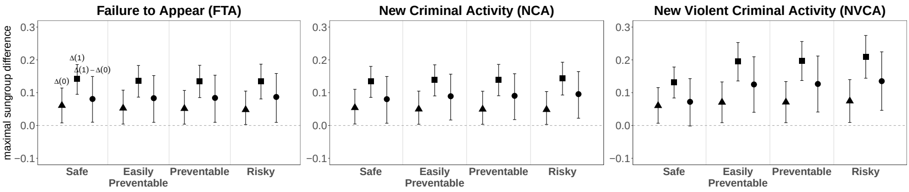
$$\Delta_r(1) - \Delta_r(0)$$

Gender and Racial Fairness

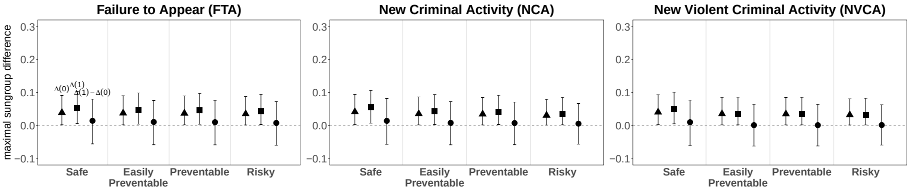


(a) Gender fairness

Gender and Racial Fairness



(a) Gender fairness



(b) Racial fairness

Concluding Remarks

- We offer a set of statistical methods for experimentally evaluating algorithm-assisted human decision making

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

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
  - improve upon the existing algorithm $\rightsquigarrow$ safe policy learning
  - dynamic interactions over time
  - multi-dimensional decision, multi-site data, multiple cases

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
  - improve upon the existing algorithm $\rightsquigarrow$ safe policy learning
  - dynamic interactions over time
  - multi-dimensional decision, multi-site data, multiple cases
- Papers available at `https://imai.fas.harvard.edu/research/`