Bringing Causality into Fairness: Application to Pretrial Public Safety Assessment

Kosuke Imai

Harvard University

Randomization, Neutrality, and Fairness
The Simons Laufer Mathematical Sciences Institute
October 24, 2023

Joint work with D. James Greiner and Zhichao Jiang

Fairness, Decision-Making, and Causality

- Fairness of decision-making must consider the impact of decision
  - admissions, hiring, insurance, lending, medical treatment, etc.
  - public policies: court decisions, government funding and programs, etc.
- We must account for how decisions influence individuals

- In contrast, the literature on algorithmic fairness focuses on prediction
- Fundamental difference between causation and association

- In this talk, I will:
  1. introduce statistical fairness criteria based on causality
  2. compare them with standard statistical fairness criteria
  3. apply them to the unique field experiment in criminal justice

Statistical Fairness Criteria

- Originally developed for assessing the fairness of prediction algorithms
- But also used for assessing the fairness of algorithmic/human decision

- Setup:
  - outcome: Y
  - prediction or decision: D
  - protected attribute (e.g., race, gender): A

- Three statistical fairness criteria:
  1. equal prediction/decision: $D \perp\!\!\!\perp A$
  2. equal accuracy: $D \perp\!\!\!\perp A \mid Y$
  3. equal calibration: $Y \perp\!\!\!\perp A \mid D$

Principal Fairness: Taking Causality into Account

- The statistical fairness criteria ignore the fact that the decision may affect the outcome
  1. fairness should address how individuals are affected by the decision
  2. observed data are contaminated (related to selective labels problem)

- Causality framework:
  - binary decision: D
  - potential outcomes: $Y(1)$ and $Y(0)$
  - different from the observed outcome: $Y = Y(D)$
  - causal effect: $Y(1) - Y(0)$
  - fundamental problem of causal inference
  - principal strata: $R = (Y(1), Y(0)) = (y_1, y_0)$

- Principal fairness:
  - accuracy: $D \perp\!\!\!\perp A \mid R$ individuals who are similarly affected by the decision should be treated similarly
  - calibration: $R \perp\!\!\!\perp A \mid D$ individuals who receive a similar decision should behave similarly

An Illustrative Example

Group A		$Y(0) = 1$	$Y(0) = 0$
		Dangerous	Backlash
$Y(1) = 1$	Detained $(D = 1)$	120	30
	Released $(D = 0)$	30	30
		Preventable	Safe
$Y(1) = 0$	Detained $(D = 1)$	70	30
	Released $(D = 0)$	70	120
Group B		$Y(0) = 1$	$Y(0) = 0$
		Dangerous	Backlash
$Y(1) = 1$	Detained $(D = 1)$	80	20
	Released $(D = 0)$	20	20
		Preventable	Safe
$Y(1) = 0$	Detained $(D = 1)$	80	40
	Released $(D = 0)$	80	160

- Satisifes principal fairness in terms of accuracy (but not calibration)
  - "Dangerous" group $(y_0 = 1, y_1 = 1)$: 80%
  - "Safe" group $(y_0 = 0, y_1 = 0)$: 20%
  - "Preventable" group $(y_0 = 1, y_1 = 0)$: 50%
  - "Backlash" group $(y_0 = 0, y_1 = 1)$: 50%

The Same Example Does Not Satisfy Statistical Fairness

	Group A		**Group B**	
	Detained	Released	Detained	Released
$Y = 1$	150	100	100	100
$Y = 0$	100	150	120	180

- This observed data are consistent with the previous example
- Statistical fairness in terms of accuracy (and calibration) is not satisfied
  - Group A: 60% ($Y = 1$), 60% ($Y = 0$)
  - Group B: 50% ($Y = 1$), 40% ($Y = 0$)

Relations between Principal Fairness and Statistical Fairness

Theorem 1

1. *If $A \perp\!\!\!\perp R$ holds, principal fairness implies all three statistical fairness criteria*
2. *If $A \perp\!\!\!\perp R$ and $Y(1) \leq Y(0)$ (i.e., monotonicity) hold, principal fairness is equivalent to the three statistical fairness criteria*

- $A \perp\!\!\!\perp R$ is the equal base rate condition with potential outcomes
- The results hold conditional on covariates
- Monotonicity eliminates the "Backlash" group in our example

Empirical Evaluation and Policy Learning

- Difficulty: principal strata are unobserved
  1. partial identification
  2. unconfoundedness assumption:

$$Y(d) \perp\!\!\!\perp D \mid \mathrm{X} \quad \text{for any } d$$

  where X is the decision variables

- Unconfoundedness is plausible if X is known and observed
- Under monotonicity and unconfoundedness, we can identify principal score: $e_r(\mathrm{X}, A) = \Pr(R = r \mid \mathrm{X}, A)$
- Policy evaluation: compute $\Pr(D = 1 \mid R, A)$
- Policy learning:
  - decision rule: $D = \delta(\mathrm{X})$
  - $\Pr(\delta(\mathrm{X}) = 1 \mid R = r, A) \; = \; \mathbb{E}\left[\frac{e_r(\mathrm{X},A)}{\mathbb{E}\{e_r(\mathrm{X},A) \mid A\}}\delta(\mathrm{X}) \mid A\right]$
  - optimal policy subject to the fairness constraint

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:

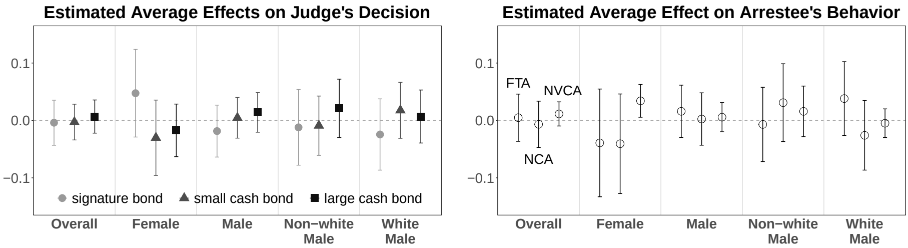
Release Recommendation - Signature bond

Conditions - Report to and comply with pretrial supervision

PSA Provision, Demographics, and Outcomes

	no PSA			PSA			
	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)

Intention-to-Treat Analysis of PSA Provision



- Insignificant average effects on a judge's decisions and arrestee's behavior

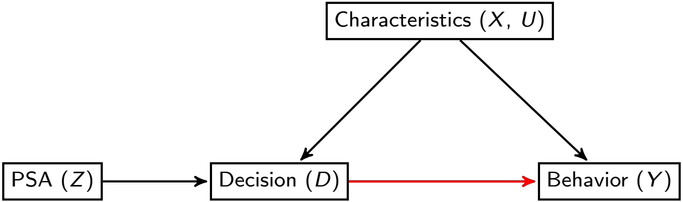
- Does PSA provision help a judge make better decisions?
- Good decision: detain risky arrestees, release safe arrestees
- Need to explore causal heterogeneity based on risk-levels

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:

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

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

Partial Identification

- 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
- Principal score

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

- Identification formula

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

an analogous formula applies to risky and safe groups

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

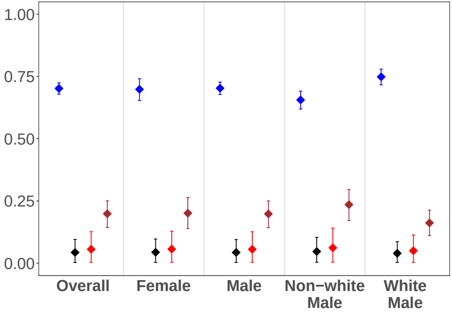
- For people with $R_i = r$
  - judges make decision $D_i \geq r \rightsquigarrow$ not commit a crime
  - judges make decision $D_i < r \rightsquigarrow$ commit a crime

- 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
- Empirical results presented below are based on parametric modeling

Empirical Results: New Criminal Activity

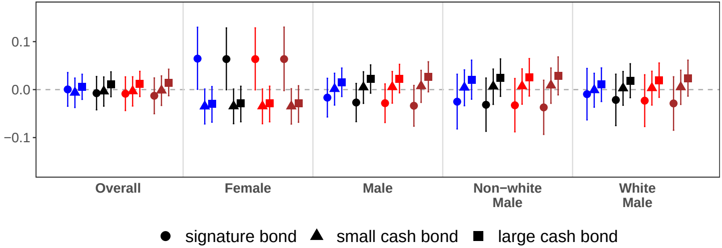


safe

easily preventable

preventable

risky



Measuring and Estimating the Degree of Fairness

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

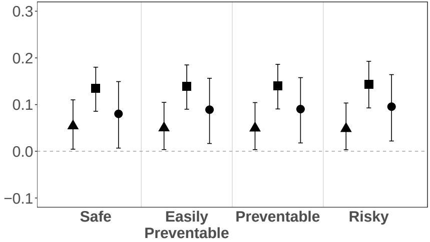
$$\Delta_r(z) = \max_{a,a',d} \left| \Pr\{D_i(z) \geq d \mid A_i = a, R_i = r\} - \Pr\{D_i(z) \geq d \mid A_i = a', R_i = r\} \right|$$

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

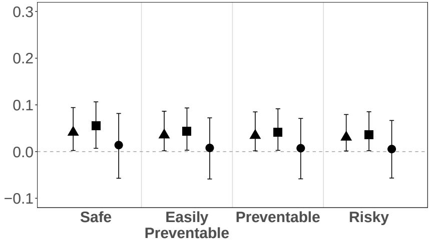
- Does the provision of PSA improve the fairness of the judge's decision?

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

(a) Gender fairness



(b) Racial fairness



▲: $\Delta_r(0)$ without PSA

■: $\Delta_r(1)$ with PSA

●: $\Delta_r(1) - \Delta_r(0)$

Concluding Remarks

- Fairness of human and algorithmic decision-making needs to be placed in the causal inference framework
- We must consider how the decision affects individuals

- Principal fairness: replace observed outcomes with potential outcomes
- Challenge: causal inference requires counterfactual
- Point identification requires untestable assumptions

- Papers: `https://imai.fas.harvard.edu/research`
  - Imai, K. and Jiang, Z. (2023). "Principal fairness for human and algorithmic decision-making." *Statistical Science*
  - Imai, K., Z. Jiang, D. J. Greiner, et al. (2023). "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." (with discussion) *Journal of the Royal Statistical Society, Series A (Statistics in Society)*