Does AI help humans make better decisions?
A methodological framework for experimental evaluation

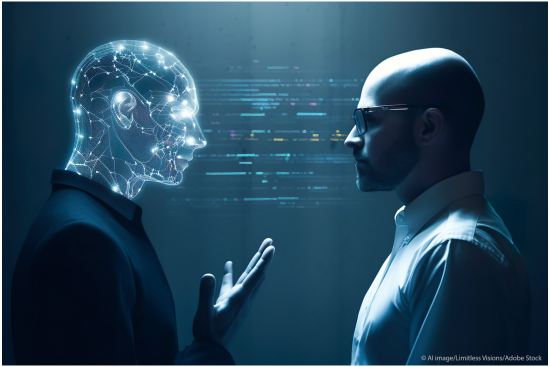
Kosuke Imai

Harvard University

Toronto Data Workshop
Faculty of Information and Department of Statistics
University of Toronto
May 3, 2024

Joint work with Eli Ben-Michael, D. James Greiner, Melody Huang, Zhichao Jiang, and Sooahn Shin

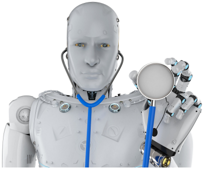
Rise of Artificial Intelligence (AI)



- Massive technological advances in recent years
- Data-driven algorithms are everywhere in our daily lives
- Generative algorithms may soon replace simple human tasks

AI-Assisted (Algorithm-Assisted) Human Decision Making

- But, humans still make many consequential decisions
- We have not yet outsourced high-stakes decisions to AI



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

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

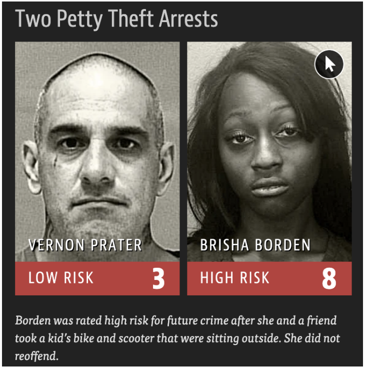
Questions and Contributions

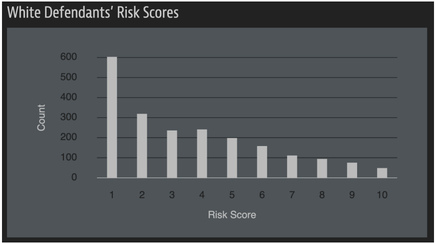
- How do AI recommendations influence human decisions?
  - Does AI help humans make more accurate decisions?
  - Does AI help humans improve the fairness of their decisions?

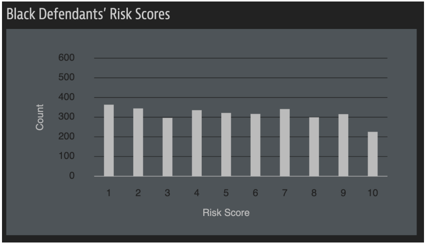
- Many have studied the accuracy and fairness of AI recommendations
  - Relatively few have researched their impacts on human decisions
  - Little is known about how AI's bias interacts with human bias

- Methodological framework for experimental evaluation
  1. experimental design: randomize human-alone vs. human+AI decisions
  2. methodology: comparison between human-alone, human+AI, AI-alone
  3. first ever field experiment: evaluating pretrial public safety assessment

Controversy over the COMPAS Score (Propublica)







Pretrial Public Safety Assessment (PSA)

- AI 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 AI 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 PSA details
  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 bail
- 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
  - we have made the data set publicly available!

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:

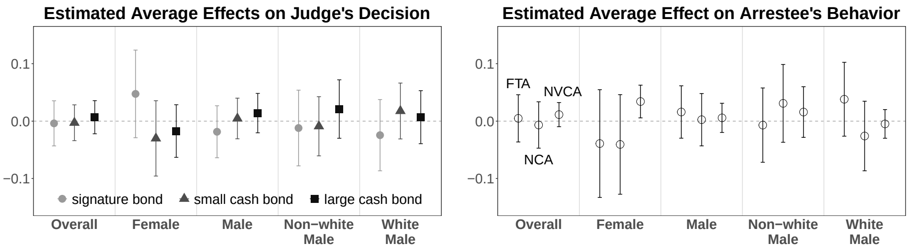
Release Recommendation - Signature bond

Conditions - Report to and comply with pretrial supervision

PSA Provision, Demographics, and Outcomes

	no PSA			PSA			
	Signature bond	Cash bail *small*	Cash bail *large*	Signature bond	Cash bail *small*	Cash bail *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 (ITT) Analysis of PSA Provision



- Mostly insignificant effects on judge's decisions (on average)
- Similar results for arrestee's behavior

- But, ITT analysis cannot answer the key question:

  Does PSA provision help judges make better decisions?

- Instead, ITT analysis asks:

  Does PSA provision influence judge's decisions?

Does the Judge Agree with AI?

Human (rows) vs. **AI** (columns)

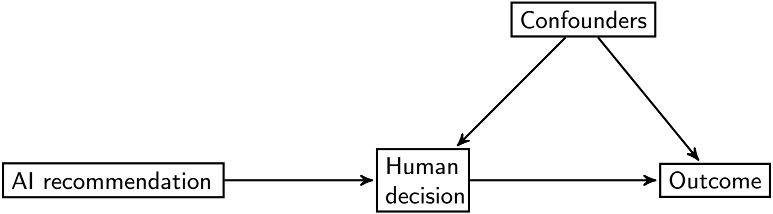
	Signature bond	Cash bail
Signature bond	54.1% (510)	20.7 (195)
Cash bail	9.4 (89)	15.8 (149)

Human+AI (rows) vs. **AI** (columns)

	Signature bond	Cash bail
Signature bond	57.3% (543)	17.1 (162)
Cash bail	7.4 (70)	18.2 (173)

Experimental Design

- Two key design features about treatment assignment:
  1. randomization: human-alone vs. human+AI
  2. single blindedness: AI recommendations affect the outcome only through human decisions
- The proposed design is widely applicable even when stakes are high



Design-based Assumptions

- Notation
  - AI recommendation provision (PSA or not): $Z_i \in \{0, 1\}$
  - Human decision (signature bond vs. cash bail): $D_i \in \{0, 1\}$
  - Observed outcome (FTA, NCA, or NVCA): $Y_i \in \{0, 1\}$
  - Potential decisions and outcomes: $D_i(z)$, $Y_i(z, D_i(z))$
- Assumptions
  1. Single-blinded treatment:

     $$Y_i(0, D_i(0)) = Y_i(1, D_i(1)) \quad \text{if} \quad D_i(0) = D_i(1) \quad \text{for all } i$$

     we can write $Y_i(z, D_i(z))$ as $Y_i(D_i(z))$
  2. Randomized treatment:

     $$Z_i \perp\!\!\!\perp \{A_i, D_i(0), D_i(1), Y_i(0), Y_i(1)\} \quad \text{for all } i$$

- These assumptions can be guaranteed by the experimental design
- Stratified randomization based on pre-treatment covariates is possible
- No other assumptions are required

Classification Ability of Decision-making System

		Decision	
		Negative ($D = 0$)	Positive ($D = 1$)
Outcome	Negative ($Y(0) = 0$)	True Negative (TN)	False Positive (FP)
	Positive ($Y(0) = 1$)	False Negative (FN)	True Positive (TP)

- Decision
  - Positive: cash bail
  - Negative: signature bond
- Outcome
  - Positive: NCA
  - Negative: no NCA

- Classification ability measures
  - False Positive (FP): unnecessary cash bail
  - False Negative (FN): signature bond followed by NCA

- Consideration of $Y(1)$ requires additional assumptions (Imai et al. JRSSA)

Classification Risk

		Decision	
		Negative ($D = 0$)	Positive ($D = 1$)
Outcome	Negative ($Y(0) = 0$)	True Negative (TN) ℓ_{00}	False Positive (FP) ℓ_{01}
	Positive ($Y(0) = 1$)	False Negative (FN) $\ell_{10} = 1$	True Positive (TP) ℓ_{11}

- Assign a (possibly asymmetric) 'loss' to each classification outcome
- Classification risk:

$$R(\ell_{01}) = \ell_{10} \cdot \mathrm{FNP} + \ell_{01} \cdot \mathrm{FPP} = q_{10} + \ell_{01} \cdot q_{01},$$

  where $q_{yd} = \Pr(Y(0) = y, D = d)$ for $y, d \in \{0, 1\}$
- Other classification ability measures:
  - misclassification rate: $R(1) = \mathrm{FNP} + \mathrm{FPP}$
  - $\mathrm{FNR} = q_{10}/(q_{10} + q_{11})$, $\mathrm{FPR} = q_{01}/(q_{00} + q_{01})$
  - false discovery rate: $\mathrm{FDR} = q_{01}/(q_{01} + q_{11})$

Comparing Human Decisions with and without AI

- Define:

$$p_{yda}(z) \;:=\; \Pr(Y(0)=y, D(z)=d, A=a)$$

- Confusion matrix:

$$\begin{aligned} C_{\text{Human}}(z) &= \begin{bmatrix} p_{000}(z)+p_{001}(z) & p_{010}(z)+p_{011}(z) \\ p_{100}(z)+p_{101}(z) & p_{110}(z)+p_{111}(z) \end{bmatrix} \\ &= \begin{bmatrix} p_{00\cdot}(z) & p_{01\cdot}(z) \\ p_{10\cdot}(z) & p_{11\cdot}(z) \end{bmatrix} \qquad \text{marginalize over AI recommendations} \end{aligned}$$

where $z=1$ is *Human+AI* and $z=0$ is *Human-alone*

- Selective labels problem: we do not observe $Y(0)$ when $D=1$
- Some elements of the confusion matrix are not identifiable

Risk Difference between Human-alone and Human+AI

- We can identify the *risk difference* between Human-alone and Human+AI systems:

$$\underbrace{\Pr(Y(0)=0 \mid Z=1)}_{p_{01\cdot}(1)+p_{00\cdot}(1)} = \underbrace{\Pr(Y(0)=0 \mid Z=0)}_{p_{01\cdot}(0)+p_{00\cdot}(0)} \text{ by randomization}$$

$$p_{01\cdot}(1) - p_{01\cdot}(0) = p_{00\cdot}(0) - p_{00\cdot}(1)$$

- Identification result:

$$\begin{aligned} & R_{\text{Human+AI}}(\ell_{01}) - R_{\text{Human}}(\ell_{01}) \\ = \; & (p_{10\cdot}(1) + \ell_{01} p_{01\cdot}(1)) - (p_{10\cdot}(0) + \ell_{01} p_{01\cdot}(0)) \\ = \; & p_{10\cdot}(1) - p_{10\cdot}(0) + \ell_{01}\,(p_{00\cdot}(0) - p_{00\cdot}(1)) \end{aligned}$$

- Hypothesis test given the relative loss ℓ_{01}:

$$H_0 : R_{\text{Human}}(\ell_{01}) \leq R_{\text{Human+AI}}(\ell_{01}),\; H_1 : R_{\text{Human}}(\ell_{01}) > R_{\text{Human+AI}}(\ell_{01})$$

- Invert this test to obtain a confidence interval on ℓ_{01}

Comparing AI Decisions with Human-alone and Human+AI

- What happens if we completely outsource decisions to AI?
- No experimental arm for AI-alone decision system

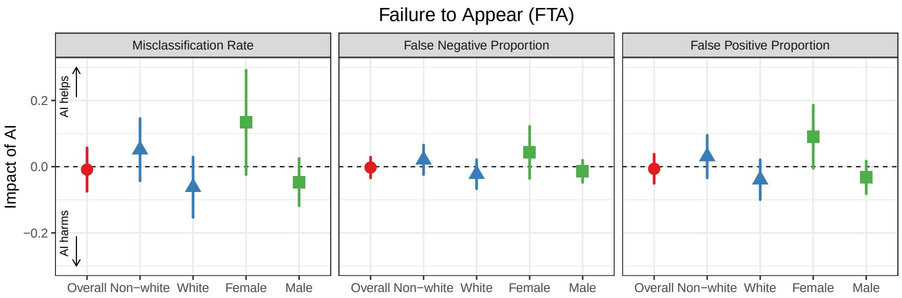
$$\begin{aligned} C_{\text{AI}} &= \begin{bmatrix} p_{000}(z) + p_{010}(z) & p_{001}(z) + p_{011}(z) \\ p_{100}(z) + p_{110}(z) & p_{101}(z) + p_{111}(z) \end{bmatrix} \\ &= \begin{bmatrix} p_{0\cdot 0}(z) & p_{0\cdot 1}(z) \\ p_{1\cdot 0}(z) & p_{1\cdot 1}(z) \end{bmatrix} \end{aligned}$$

- Bound the risk differences, $R_{\text{AI}}(\ell_{01}) - R_{\text{Human}}(\ell_{01})$ and $R_{\text{AI}}(\ell_{01}) - R_{\text{Human+AI}}(\ell_{01})$, using:

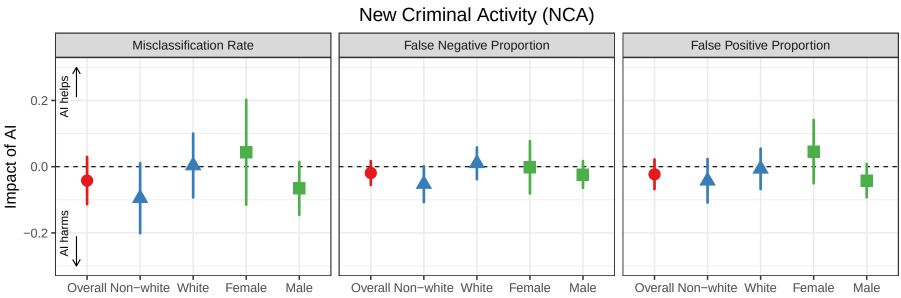
$$\begin{aligned} p_{y1a}(z) &= \underbrace{\Pr(Y(0) = y \mid D(z) = 1, Z = z, A = a)}_{\in[0,1]} \\ &\quad \times P(D(z) = 1 \mid A = a, Z = z) \cdot \Pr(A = a) \\ &\in [0,\ \Pr(D = 1 \mid A = a, Z = z) \Pr(A = a)] \end{aligned}$$

- Sharp bounds are more complex and only slightly tighter

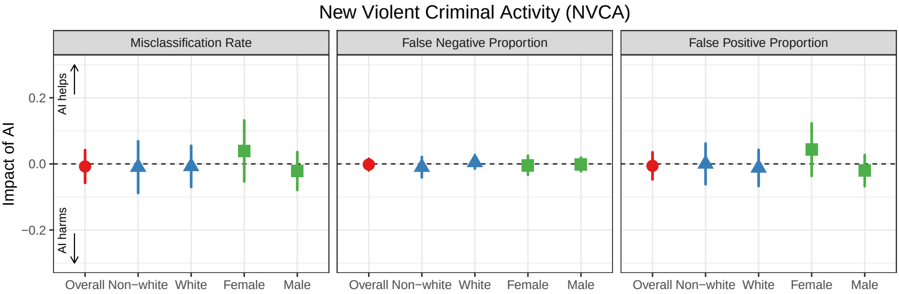
AI Recommendations Do Not Improve Human Decisions



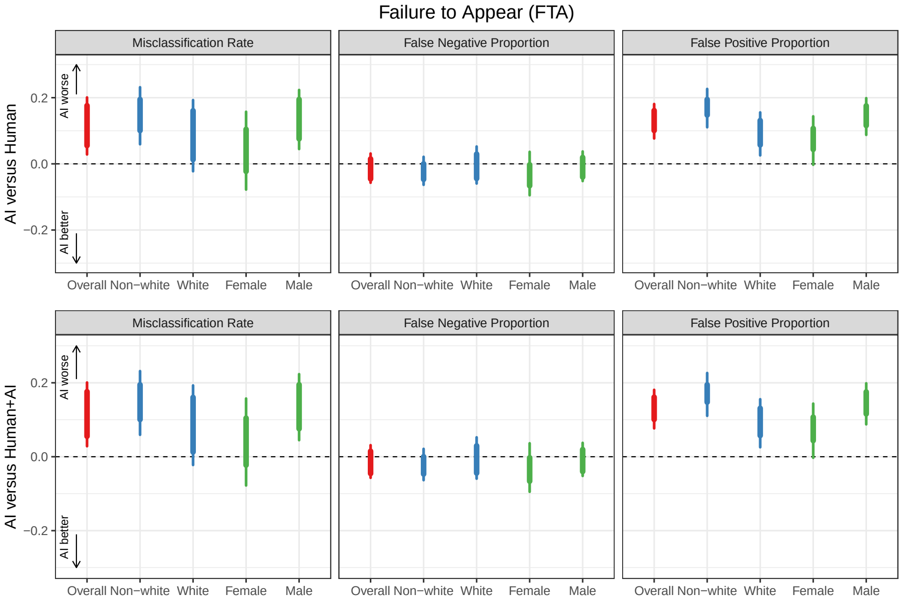
AI Recommendations Do Not Improve Human Decisions



AI Recommendations Do Not Improve Human Decisions

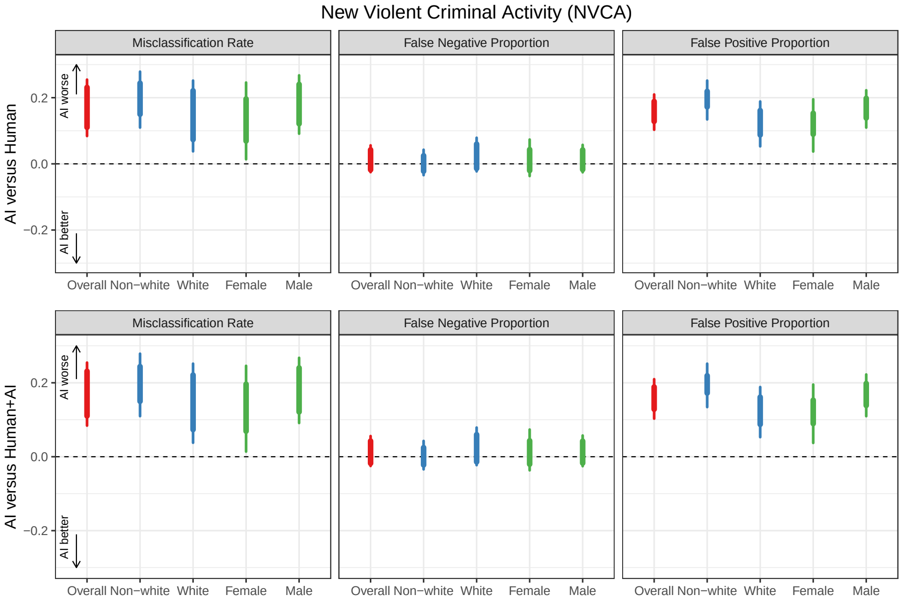


AI-Alone Decisions Perform Worse than Human Decisions

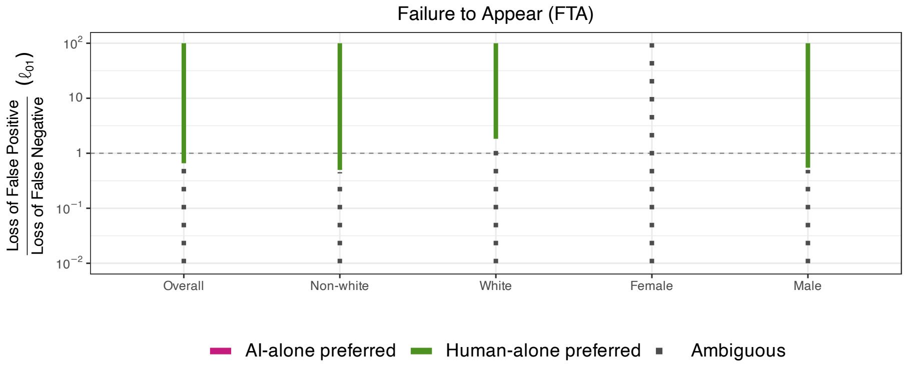


AI-Alone Decisions Perform Worse than Human Decisions

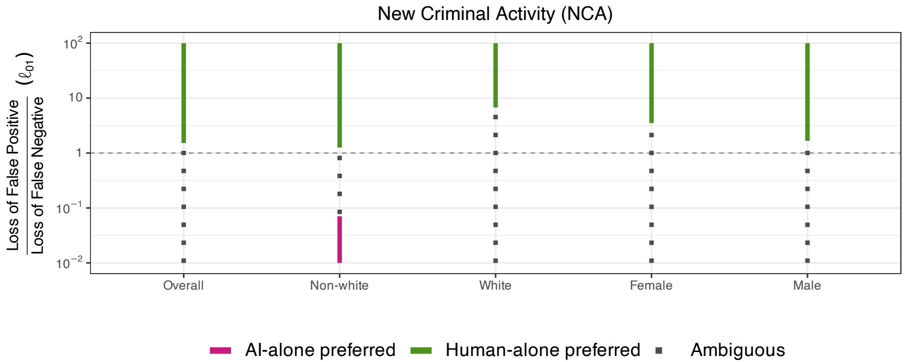
AI-Alone Decisions Perform Worse than Human Decisions



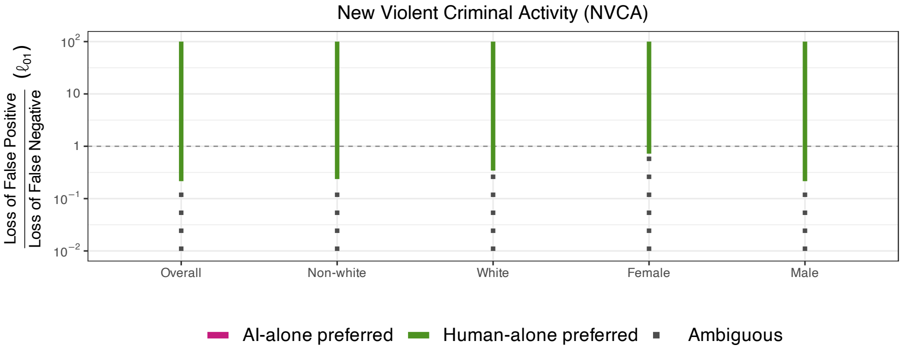
Human-Alone System is Preferred over AI-Alone System when the Cost of False Positive is High



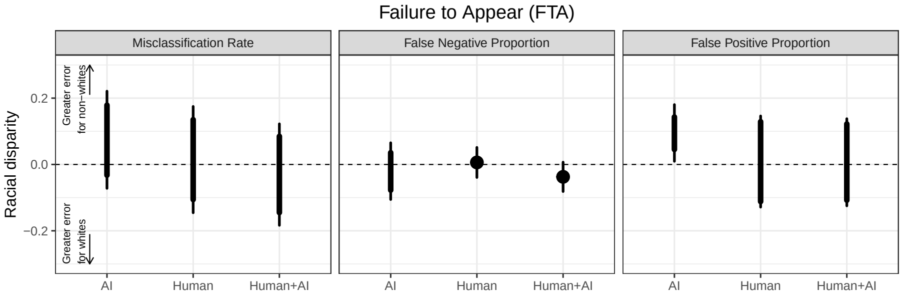
Human-Alone System is Preferred over AI-Alone System when the Cost of False Positive is High



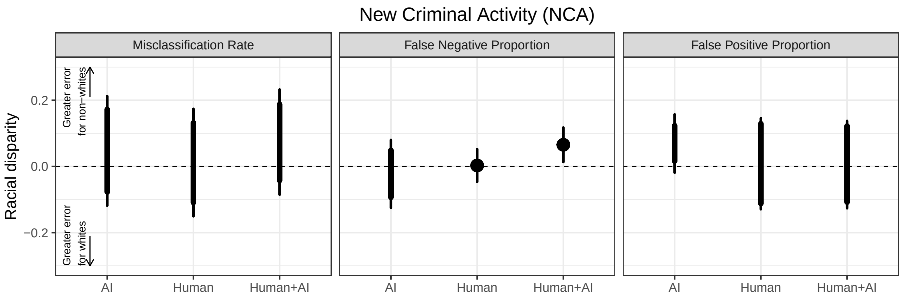
Human-Alone System is Preferred over AI-Alone System when the Cost of False Positive is High



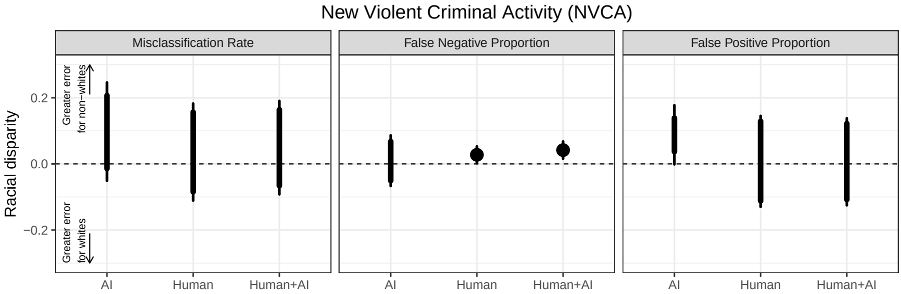
AI-Alone System Has More False Positives for Non-whites



AI-Alone System Has More False Positives for Non-whites



AI-Alone System Has More False Positives for Non-whites



Concluding Remarks

- We propose a methodological framework for experimentally evaluating the three decision-making systems:
  1. Human-alone
  2. Human+AI
  3. AI-alone

- The proposed methodological framework is widely applicable
  - single-blinded treatment assignment is easy to implement
  - do not require AI-alone treatment condition
  - no additional assumption is required
  - open-source R software package aihuman is available

- We conducted and analyzed an RCT that evaluates the pretrial risk assessment instrument (PSA-DMF sytem):
  1. AI recommendations have little impacts on human decisions
  2. AI decisions perform worse than human decisions

PSA Scoring Rule

Risk factor		FTA	NCA	NVCA
Current violent offense	$>$ 20 years old			2
	$\leq$ 20 years old			3
Pending charge at time of arrest		1	3	1
Prior conviction	misdemeanor or felony	1	1	1
	misdemeanor and felony	1	2	1
Prior violent conviction	1 or 2		1	1
	3 or more		2	2
Prior sentence to incarceration			2	
Prior FTA in past 2 years	only 1	2	1	
	2 or more	4	2	
Prior FTA older than 2 years		1		
Age	22 years or younger		2	

- FTA: $\{0 \to 1, 1 \to 2, 2 \to 3, (3,4) \to 4, (5,6) \to 5, 7 \to 6\}$
- NCA: $\{0 \to 1, (1,2) \to 2, (3,4) \to 3, (5,6) \to 4, (7,8) \to 5,$ $(9,10,11,12,13) \to 6\}$
- NVCA: $\{(0,1,2,3) \to 0, (4,5,6,7) \to 1\}$

Decision Making Framework (DMF)