# Does AI help humans make better decisions?
## A statistical evaluation framework for experimental and observational studies

Kosuke Imai

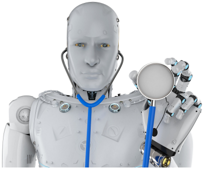
Harvard University

Department of Economics Seminar
University of Fribourg, Switzerland
October 8, 2024

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

# AI-assisted (Algorithm-assisted) human decision making

- AI and data-driven algorithms are everywhere in our daily lives
- 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.

# Key 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

- A statistical evaluation framework for AI recommendations
  1. experimental studies: randomize human-alone vs. human+AI decisions
  2. observational studies: also applicable under unconfoundedness
  3. methodology:
     - compare human-alone, human+AI, and AI-alone
     - optimally combine human decisions with AI recommendations
  4. first ever field experiment: evaluating pretrial public safety assessment

# 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)
- Judges may have additional information we do not observe
- 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
  - used in more than 25 states

# 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
  - PSA is calculated for each case using a computer system
  - provision of PSA is randomized across cases
  - 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]
**DOB:** [redacted]
**Arrest Date:** 03/25/2017

**Spillman Name Number:** [redacted]
**Gender:** Male
**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) |

# Does the judge agree with PSA?

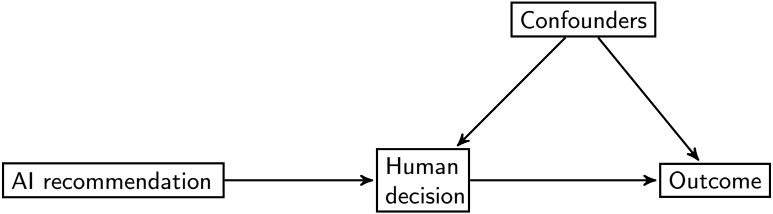
| | | PSA | |
|---|---|---|---|
| | | Signature bond | Cash bail |
| Human | Signature bond | 54.1% (510) | 20.7 (195) |
| | Cash bail | 9.4 (89) | 15.8 (149) |

| | | PSA | |
|---|---|---|---|
| | | Signature bond | Cash bail |
| Human+PSA | Signature bond | 57.3% (543) | 17.1 (162) |
| | Cash bail | 7.4 (70) | 18.2 (173) |

- PSA statistically significantly influence the judge's decision
- But how?

# Experimental design

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



# Required 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. Unconfounded treatment:

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

  3. Overlap: $0 < \Pr(Z_i = 1 \mid X_i = x) < 1$ for all $x$
- These assumptions can be guaranteed by the experimental design
- 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) |

- (Generic) Decision $D^*$
  - Positive: cash bail
  - Negative: signature bond
- Outcome under release $Y(0)$
  - Positive: NCA
  - Negative: no NCA

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

- We focus on $Y(0)$ and ignore $Y(1)$

# Classification risk

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

- Assign a (possibly asymmetric) 'loss' to each classification outcome
- Classification risk of decision-making system $D^*$

$$R(\ell_{01}; D^*) \ := \ \underbrace{\ell_{10}}_{=1} \cdot \underbrace{p_{10}(D^*)}_{\text{FNP}} + \ell_{01} \cdot \underbrace{p_{01}(D^*)}_{\text{FPP}},$$

  where $p_{yd}(D^*) = \Pr(Y(0) = y, D^* = d)$ for $y, d \in \{0, 1\}$
- misclassification rate: $R(1; D^*) = \text{FNP} + \text{FPP}$

# Comparing human decisions with and without AI

- Risk difference:

$$
\begin{aligned}
& R_{\text{human+AI}}(\ell_{01}) - R_{\text{human}}(\ell_{01}) \\
= & \{p_{10}(D(1)) - p_{10}(D(0))\} + \ell_{01}\{p_{01}(D(1)) - p_{01}(D(0))\}
\end{aligned}
$$

- Selective labels problem: we do not observe $Y(0)$ when $D = 1$
- FNP is identifiable but FPP is unidentified

- The difference of FPP is identifiable
  - by randomization
    $\Pr(Y(0) = 0 \mid Z = 1, X = x) = \Pr(Y(0) = 0 \mid Z = 0, X = x)$
  - by law of total probability

$$
\begin{aligned}
& p_{01}(D(1) \mid X = x) + p_{00}(D(1) \mid X = x) \\
= & p_{01}(D(0) \mid X = x) + p_{00}(D(0) \mid X = x)
\end{aligned}
$$

# Doubly robust estimation

- Identification formula:

$$\begin{aligned}
&R_{\text{human+AI}}(\ell_{01}) - R_{\text{human}}(\ell_{01}) \\
= &\,\mathbb{E}\left[\Pr(Y=1, D=0 \mid Z=1, X) - \Pr(Y=1, D=0 \mid Z=0, X)\right. \\
&\left. -\ell_{01}\left\{\Pr(Y=0, D=0 \mid Z=1, X) - \Pr(Y=0, D=0 \mid Z=0, X)\right\}\right]
\end{aligned}$$

- Compound outcome: $W_i := Y_i(1-D_i) - \ell_{01}(1-Y_i)(1-D_i)$
- Three models:
  1. propensity score: $e(z,x) := \Pr(Z=z \mid X=x)$
  2. decision model: $m^D(z,x) := \Pr(D=1 \mid Z=z, X=x)$
  3. outcome model: $m^Y(z,x) := \Pr(Y=1 \mid D=0, Z=z, X=x)$

- AIPW estimator:

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\{\widehat{\varphi}_1(Z_i, X_i, D_i, Y_i; \ell_{01}) - \widehat{\varphi}_0(Z_i, X_i, D_i, Y_i; \ell_{01})\}$$

where $\widehat{\varphi}_z(Z, X, D, Y; \ell_{01})$ is the (uncentered) influence function:

$$\begin{aligned}
&\widehat{\varphi}_z(Z, X, D, Y; \ell_{01}) \\
:= &\left(1 - \hat{m}^D(z, X)\right)\left\{(1+\ell_{01})\hat{m}^Y(z, X) - \ell_{01}\right\} \\
&\quad + (1+\ell_{01})\frac{\mathbb{1}\{Z = z\}(1-D)}{\hat{e}(z, X)}\left(Y - \hat{m}^Y(z, X)\right) \\
&\qquad - \left\{(1+\ell_{01})\hat{m}^Y(z, X) - \ell_{01}\right\}\frac{\mathbb{1}\{Z = z\}}{\hat{e}(z, X)}\left(D - \hat{m}^D(z, X)\right)
\end{aligned}$$

- Properties:
  - asymptotic normality
  - double robustness: (outcome model + decision model) $\times$ propensity score model

# When do you prefer human-alone vs. human+AI?

- Hypothesis test given the relative loss $\ell_{01}$:

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

- Invert this test to obtain a confidence interval on $\ell_{01}$
  1. Reject $H_0$: prefer Human+AI over Human-alone
  2. Reject $H_1$: prefer Human-alone over Human+AI
  3. Fail to reject either hypothesis: statistically ambiguous

# 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

$$R_{\mathrm{AI}}(\ell_{01}) \ := \ R(\ell_{01}; A) \ = \ p_{10}(A) + \ell_{01} p_{01}(A)$$

where

$$p_{ya}(A) = \Pr(Y(0) = y, A = a, D = 1) + \Pr(Y(0) = y, A = a, D = 0)$$

- Derive the sharp bound of risk difference: e.g., $R_{\mathrm{AI}}(\ell_{01}) - R_{\mathrm{Human}}(\ell_{01})$
- The bound width depends on the agreement between Human and AI:

$$(1+\ell_{01})\mathbb{E}\left\{\Pr(A=0 \mid X) - \max_{z'} \Pr(Y=1, D=0, A=0 \mid Z=z', X) - \max_{z'} \Pr(Y=0, D=0, A=0 \mid Z=z', X)\right\}$$

- Applicable to any generic AI or any other decision system

# Doubly robust estimation

- Estimation of bounds is complex
  1. data-driven choice of $z'$
  2. estimation of the bounds given the optimal choice of $z'$
- Decision and outcome models:
  1. $m^D(z, x, a) := \Pr(D = 1 \mid Z = z, X = x, A = a)$
  2. $m^Y(z, x, a) := \Pr(Y = 1 \mid D = 0, Z = z, X = x, A = a)$
- Nuisance classifier for the lower bound:

$$g_{L_z}(x) = \mathbb{1}\{(1-m^D(1-z, x, 0))m^Y(1-z, x, 0) \geq (1-m^D(z, x, 0))m^Y(z, x, 0)\}$$

  - assume that this nuisance classifier is well separated
  - plug-in estimation
- Compound outcomes: $Y(1-D)(1-A)$, $(1-Y)(1-D)(1-A)$, $(1-A)D$, and $A(1-D)$
- AIPW: asymptotic normality, double-robustness

# When do you prefer Ai-alone vs. Human-alone?

- Same hypothesis testing framework as before:

$$H_0 : R_{\mathrm{AI}}(\ell_{01}) \leq R_{\mathrm{Human}}(\ell_{01}),$$
$$H_1 : R_{\mathrm{AI}}(\ell_{01}) > R_{\mathrm{Human}}(\ell_{01}).$$

- Due to partial identification, we instead test
  1. $H_{L0} : L_0 \leq 0$ vs. $H_{L1} : L_0 > 0$
  2. $H_{U0} : U_0 \geq 0$ vs. $H_{U1} : U_0 < 0$

- As before, we invert these hypothesis tests
  1. Rejecting $H_{L0}$ implies Human is preferred over AI
  2. Rejecting $H_{U0}$ implies AI is preferred over Human
  3. Ambiguous otherwise

# Learning when to provide AI recommendations

- Policy: $\pi : \mathcal{X} \to \{0, 1\}$, provide AI recommendation or not
- Optimal policy:

$$\pi^*_{\text{rec}} \in \operatorname*{argmin}_{\pi \in \Pi} R_{\text{rec}}(\ell_{01}; \pi)$$

where

$$\begin{aligned} R_{\text{rec}}(\ell_{01}; \pi) \; &:= p_{10}(D(\pi(X))) + \ell_{01} p_{01}(D(\pi(X))) \\ &= R_{\text{human}}(\ell_{01}) + \mathbb{E}\left[\pi(X)\left\{p_{10}(D(1) \mid X) - p_{10}(D(0) \mid X)\right.\right. \\ &\qquad \left.\left. -\ell_{01} \cdot (p_{00}(D(1) \mid X) - p_{00}(D(0) \mid X))\right\}\right] \end{aligned}$$

- Empirical risk minimization using the doubly robust score

# Learning when to follow AI recommendations

- Optimally following the AI recommendations:

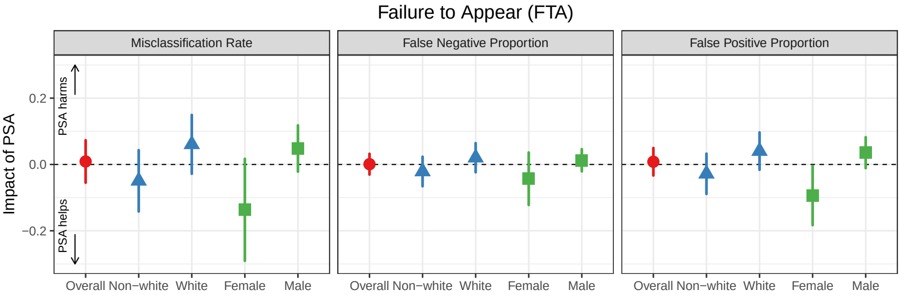
$$\pi^*_{\text{dec}} \in \underset{\pi \in \Pi}{\text{argmin}}\ \mathbb{E}[\pi(X) U_0(X)],$$

where

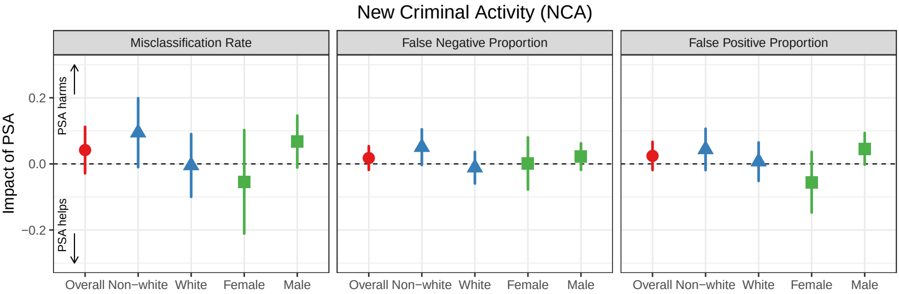
$$\begin{aligned} R_{\text{dec}}(\ell_{01}; \pi) &:= p_{10}(\widetilde{D}) + \ell_{01} p_{01}(\widetilde{D}) \\ &= R_{\text{human}}(\ell_{01}) + \mathbb{E}\left[\pi(X)\left\{p_{10}(A \mid X) - p_{10}(D(0) \mid X)\right.\right. \\ &\qquad \left.\left. + \ell_{01} \cdot (p_{01}(A \mid X) - p_{01}(D(0) \mid X))\right\}\right] \end{aligned}$$

- Use the partial identification and doubly-robust score to optimize the empirical worst-case risk (upper bound) $\rightsquigarrow$ safe policy learning

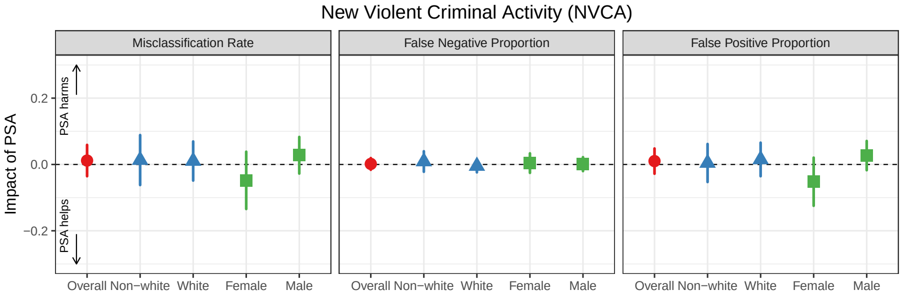
# PSA recommendations do not improve human decisions



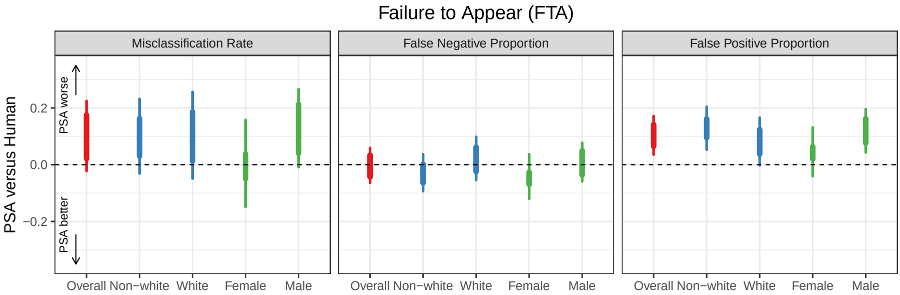
# PSA recommendations do not improve human decisions



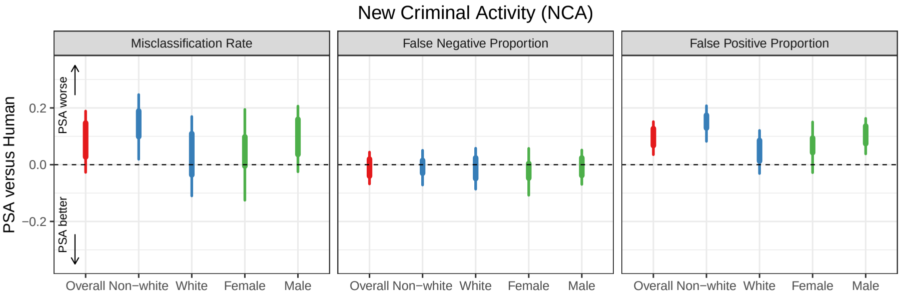
# PSA recommendations do not improve human decisions



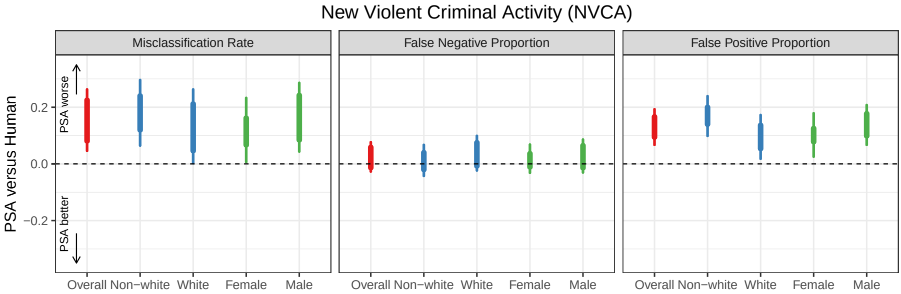
# PSA-alone decisions are less accurate than human decisions



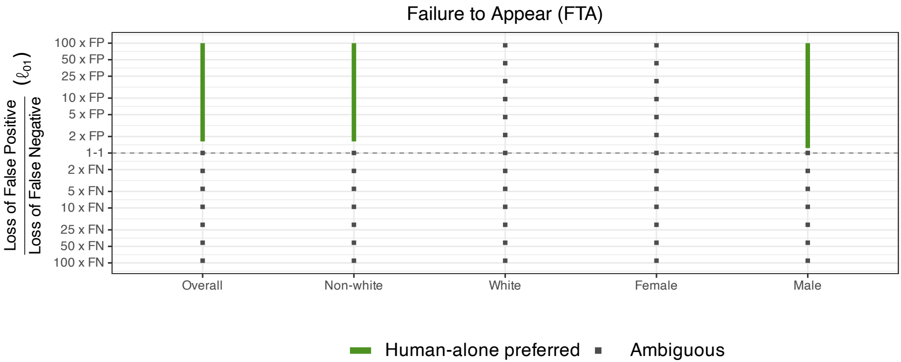
# PSA-alone decisions are less accurate than human decisions



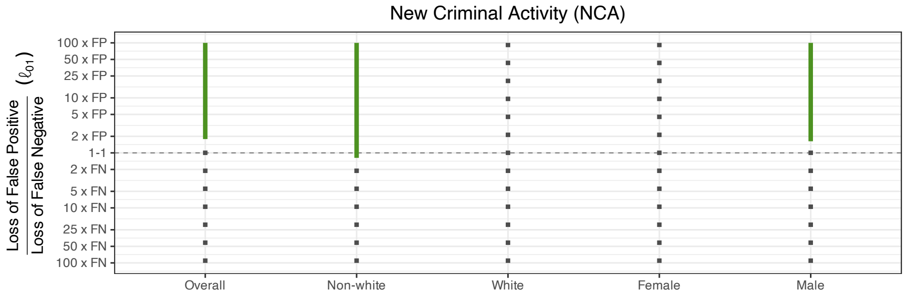
# PSA-alone decisions are less accurate than human decisions



# Human-alone system is preferred over PSA-alone system when the cost of false positive is high

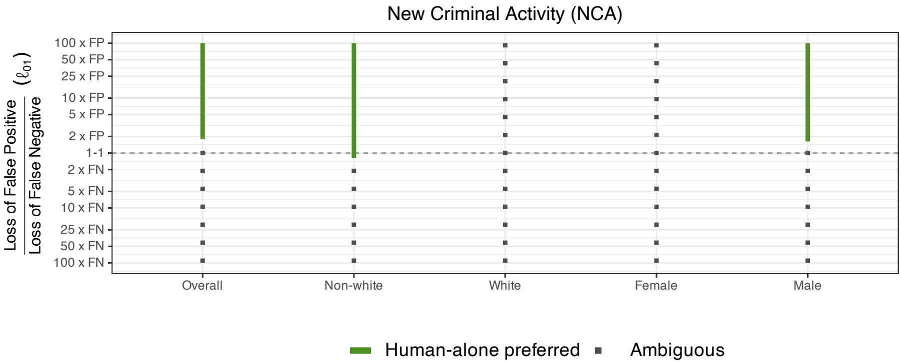


# Human-alone system is preferred over AI-alone system when the cost of false positive is high

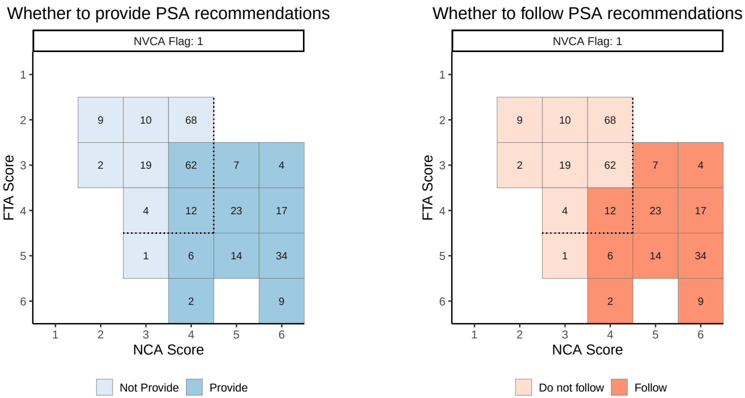



Overall Non-white White Female M

# Human-alone system is preferred over AI-alone system when the cost of false positive is high



# Optimally combining PSA recommendations with human decisions



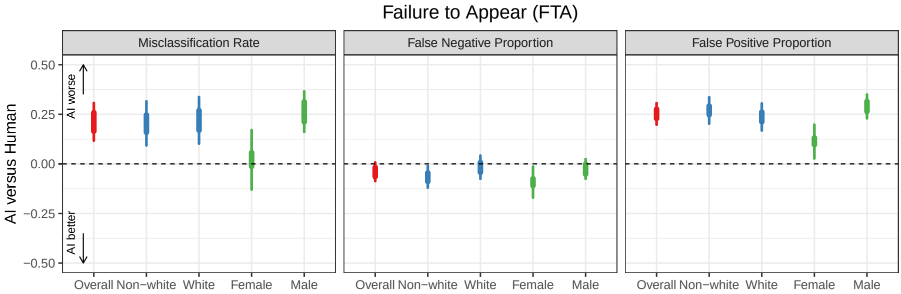

- PSA is useful only in cases with extreme recommendations

# PSA is not an AI. What about the Real AI?

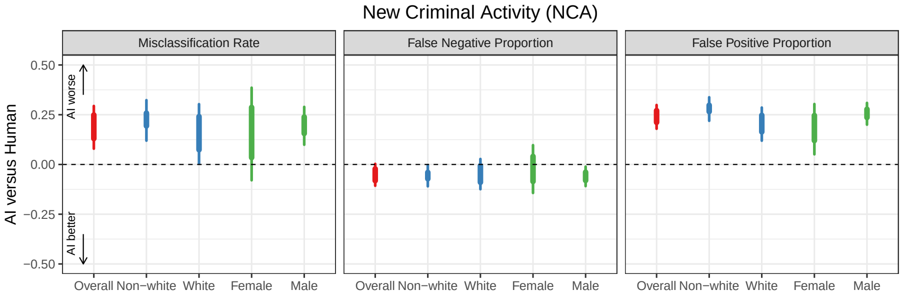
*You are a judge in Dane County, Madison, Wisconsin and are asked to decide whether or not an arrestee should be released on their own recognizance or be required to post a cash bail. If you think the risk of unnecessary incarceration is too high, then the arrestee should receive own recognizance release. On the other hand, you should assign cash bail if the following risks are too high: the risk of failure to appear at subsequent court dates, the risk of engaging in new criminal activity, and the risk of engaging in new violent criminal activity. You are provided with the following 12 characteristics about an arrestee:* ***[description of PSA inputs]****.*

*This arrestee has the following characteristics:* ***[arrestee's PSA inputs]****. Should this arrestee be released on their own recognizance or given cash bail? Please provide your answer in binary form (0 for released on their own recognizance and 1 for cash bail), followed by a detailed explanation of your decision.*

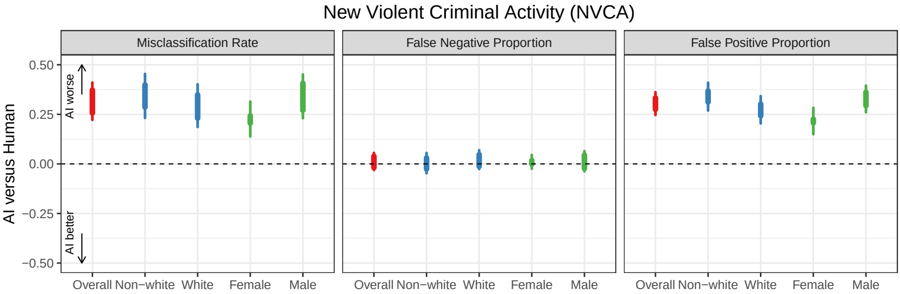
# AI-alone decisions are less accurate than human decisions



# AI-alone decisions are less accurate than human decisions



# AI-alone decisions are less accurate than human decisions



# Concluding remarks

- We propose a methodological framework for evaluating 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
  - unconfoundedness + overlap enable RCT and observational studies
  - do not require AI-alone treatment condition
  - no additional assumption is required
  - open-source R software package aihuman is available

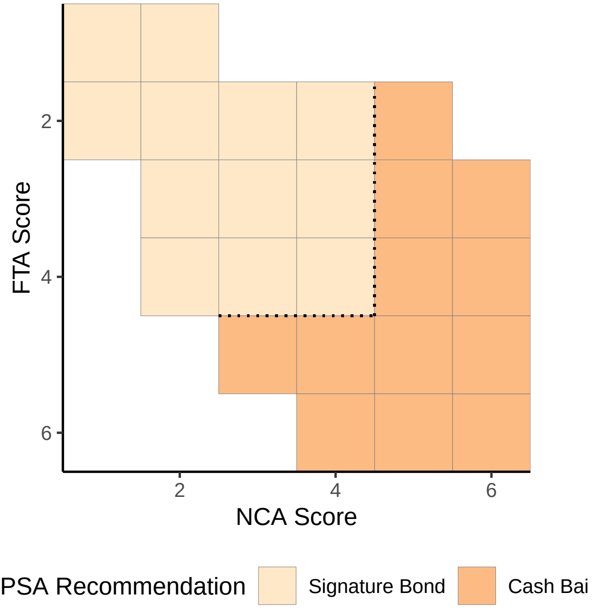
- We conducted and analyzed an RCT that evaluates the pretrial risk assessment instrument (PSA-DMF sytem):
  1. PSA recommendations do not improve human decisions
  2. Only extreme PSA recommendations are useful
  3. Both PSA and AI decisions perform worse than human decisions

# 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)




back