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

Kosuke Imai

Harvard University

Interdisciplinary Seminar at Department of Political Science
Trinity College Dublin
March 31, 2025

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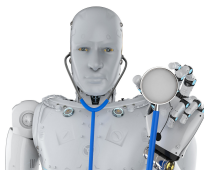
  - this is true even when human decisions can be suboptimal
  - we may want to hold *someone*, rather than *something*, accountable

- 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
  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
  - we have made the midline data set publicly available!

Name: █████████   **Spillman Name Number:** ██████

DOB: ████████   **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

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
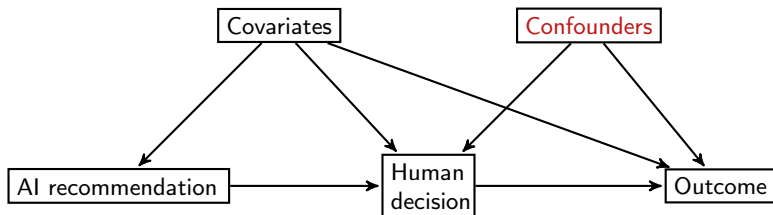
Required assumptions

- Notation
  - AI recommendation provision (PSA or not): $Z_i \in \{0, 1\}$
  - Human decision (signature bond vs. cash bail): $D_i \in \{0, 1\}$
  - AI recommendation: $A_i \in \{0, 1\}$
  - 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: for all i and d

  $$Y_i(0, d) = Y_i(1, d)$$

  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) 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 of decision-making system D^*

$$R(\ell_{01}; D^*) := \mathbb{E}[\ell_{10} \cdot \mathsf{FN} + \ell_{01} \cdot \mathsf{FP}]$$
$$= \underbrace{\ell_{10}}_{=1} \cdot \underbrace{p_{10}(D^*)}_{\mathsf{FNP}} + \ell_{01} \cdot \underbrace{p_{01}(D^*)}_{\mathsf{FPP}},$$

where $p_{yd}(D^*) = \Pr(Y(0) = y, D^* = d)$ for $y, d \in \{0, 1\}$

- misclassification rate: $R(1; D^*) = \mathsf{FNP} + \mathsf{FPP}$

Comparing human decisions with and without AI

- Risk difference:

$$R_{\text{human+AI}}(\ell_{01}) - R_{\text{human}}(\ell_{01})$$
$$= \underbrace{\{p_{10}(D(1)) - p_{10}(D(0))\}}_{\text{FNP difference}} + \ell_{01} \underbrace{\{p_{01}(D(1)) - p_{01}(D(0))\}}_{\text{FPP difference}}$$

- Selective labels problem: we do not observe $Y(0)$ when $D = 1$
- FNP is identifiable but FPP is unidentified
- But, the FPP *difference* 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
    $$p_{01}(D(1) \mid X = x) + p_{00}(D(1) \mid X = x)$$
    $$= p_{01}(D(0) \mid X = x) + p_{00}(D(0) \mid X = x)$$
    $$\iff \text{FPP difference} = -\text{TNP difference}$$

- Double machine learning

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

- 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}
  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

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

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., $R_{\mathsf{AI}}(\ell_{01}) - R_{\mathsf{Human}}(\ell_{01})$
- The bound width depends on the agreement between Human and AI

- Applicable to any generic AI or any other decision system
- Double machine learning

When do you prefer AI-alone vs. Human-alone?

- Same hypothesis testing framework as before:

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

- 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:

$$\pi_{\text{rec}}^* \in \underset{\pi \in \Pi}{\operatorname{argmin}} \underbrace{p_{10}(D(\pi(X))) + \ell_{01} p_{01}(D(\pi(X)))}_{=R_{\text{rec}}(\ell_{01}; \pi)}$$

- Empirical risk minimization using the doubly robust score

Learning when to follow AI recommendations

- Optimally following the AI recommendations (when we know the AI-alone system is better than the human decision-maker):

$$\pi^*_{\mathsf{dec}} \in \underset{\pi \in \Pi}{\arg\min} \ \underbrace{p_{10}(\widetilde{D}(\pi(X))) + \ell_{01}p_{01}(\widetilde{D}(\pi(X)))}_{R_{\mathsf{dec}}(\ell_{01};\pi)},$$

where $\widetilde{D}(\pi(X)) = A\pi(X) + D(0)(1 - \pi(X))$

- 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
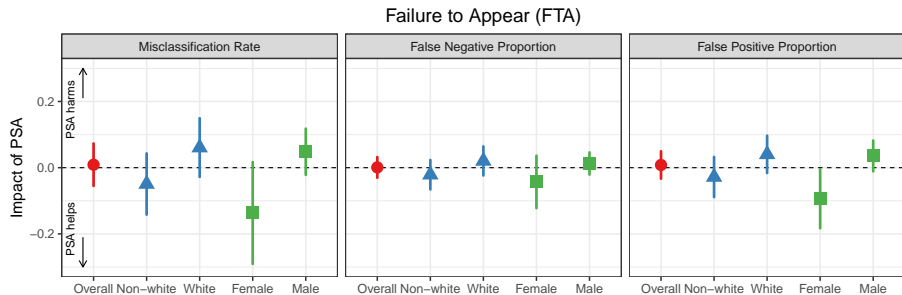
Failure to Appear (FTA)

PSA recommendations do not improve human decisions

New Criminal Activity (NCA)
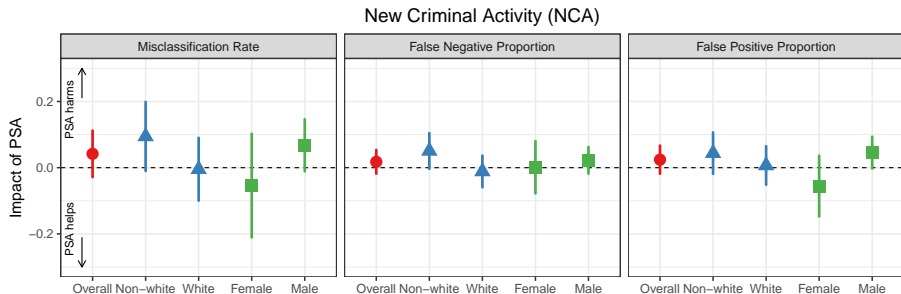
PSA recommendations do not improve human decisions

New Violent Criminal Activity (NVCA)

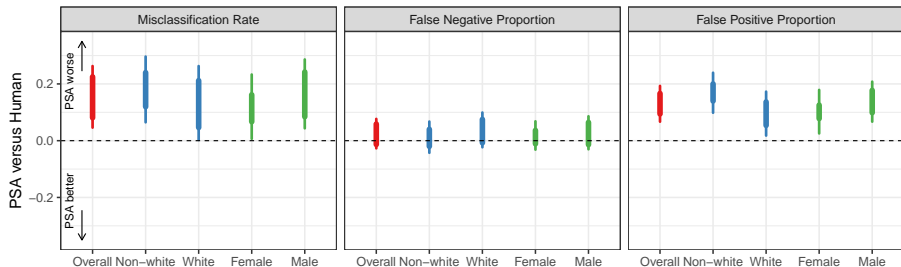# PSA-alone decisions are less accurate than human decisions
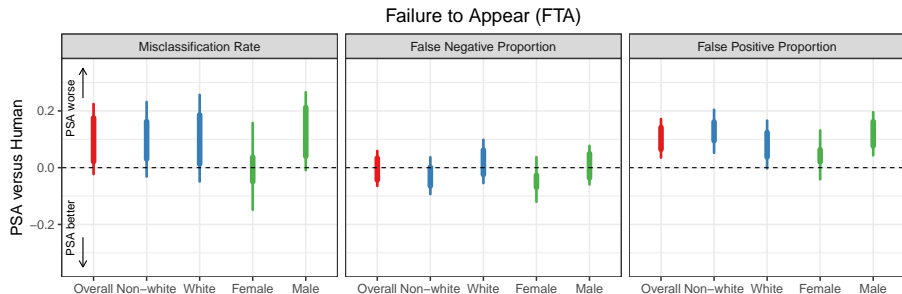
Failure to Appear (FTA)

PSA-alone decisions are less accurate than human decisions

New Criminal Activity (NCA)
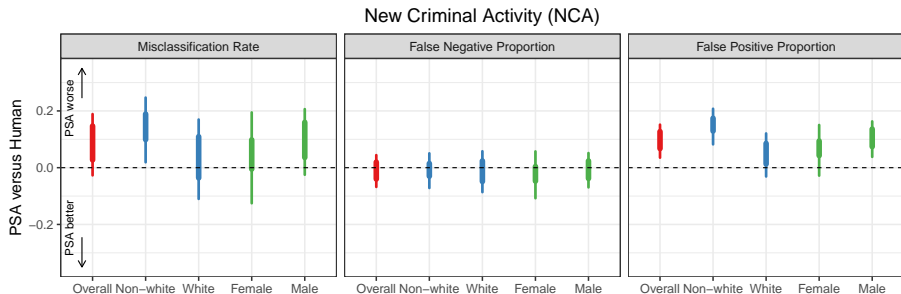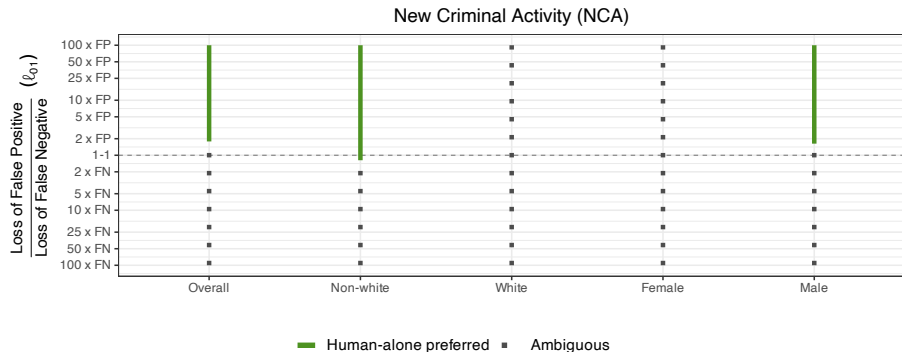
PSA-alone decisions are less accurate than human decisions

New Violent Criminal Activity (NVCA)

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

Failure to Appear (FTA)

Loss of False Positive / Loss of False Negative (ℓ_{01})

Legend: ■ Human-alone preferred ■ Ambiguous

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

New Criminal Activity (NCA)

Human-alone preferred   Ambiguous

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

New Violent Criminal Activity (NVCA)

Legend: ■ Human-alone preferred ■ Ambiguous
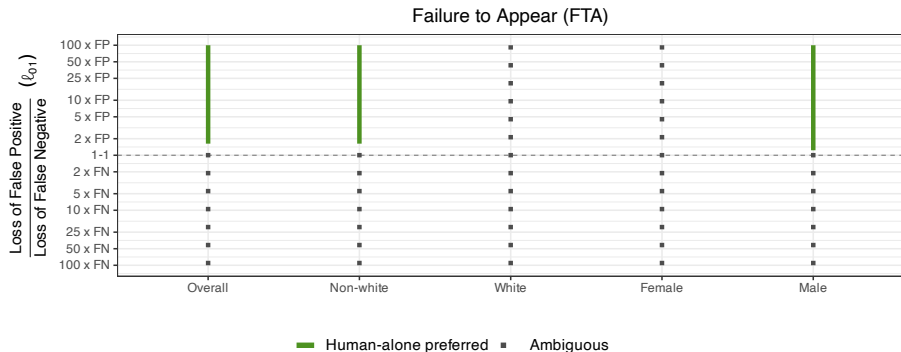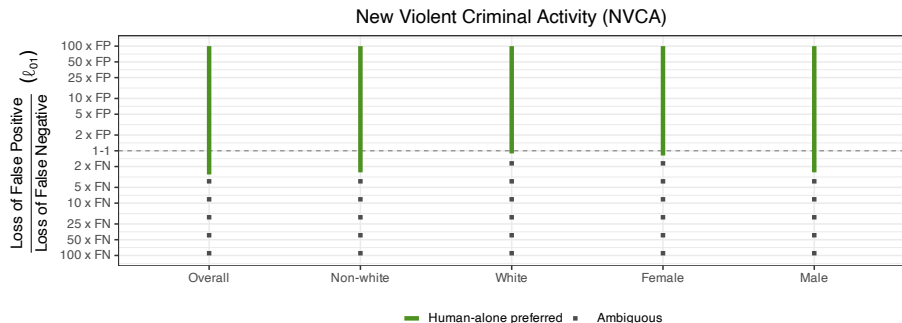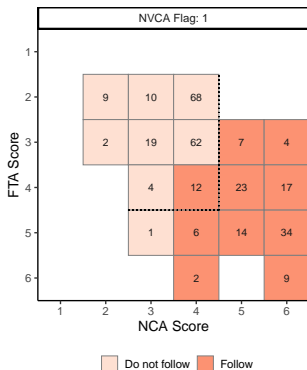
Optimally combining PSA with human decisions

Whether to provide PSA recommendations
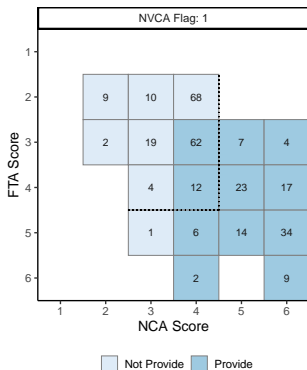
Whether to follow PSA recommendations

- 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.*
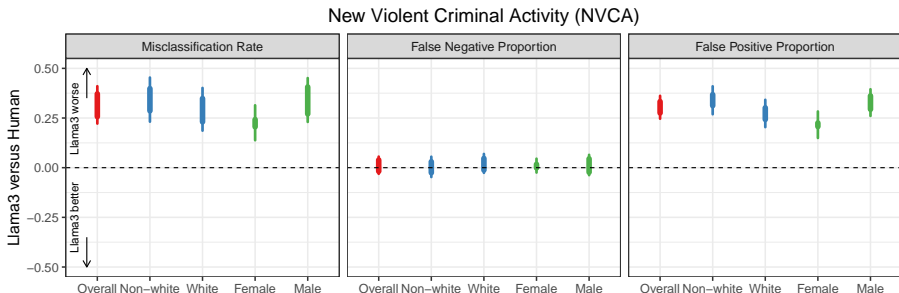
AI-alone decisions are less accurate than human decisions
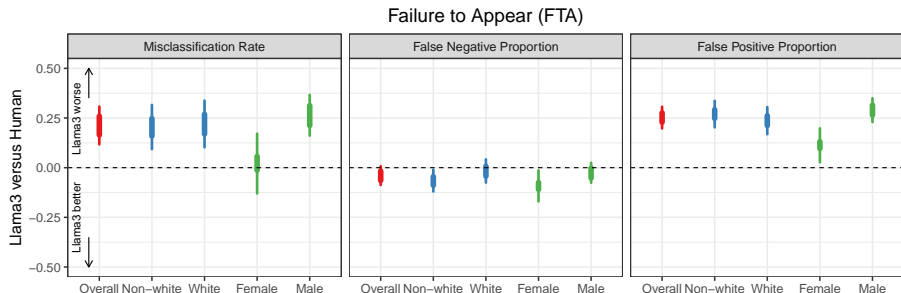
Failure to Appear (FTA)

AI-alone decisions are less accurate than human decisions

New Criminal Activity (NCA)

AI-alone decisions are less accurate than human decisions

New Violent Criminal Activity (NVCA)
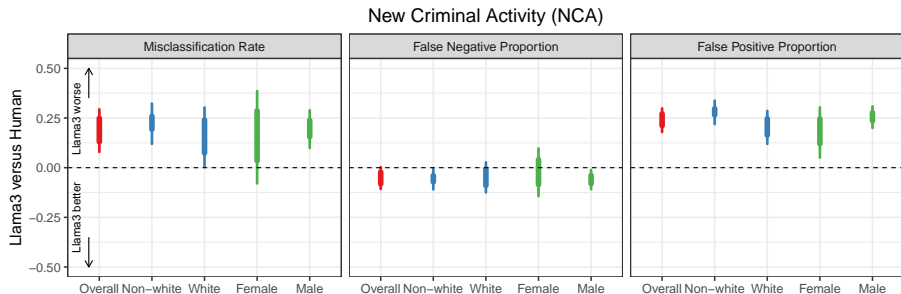
Concluding remarks

- We propose a methodological framework for evaluating three decision-making systems: Human-alone, Human+AI, AI-alone

- The proposed methodological framework is widely applicable
  - single-blinded treatment assignment is easy to implement
  - applicable to RCT and observational studies
  - do not require AI-alone treatment condition
  - no assumption other than unconfoundedness and overlap 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

- Ongoing extensions: multiple decisions, joint potential outcomes, dynamic decisions