Does AI help humans make better decisions?

A methodological framework for experimental evaluation

Kosuke Imai

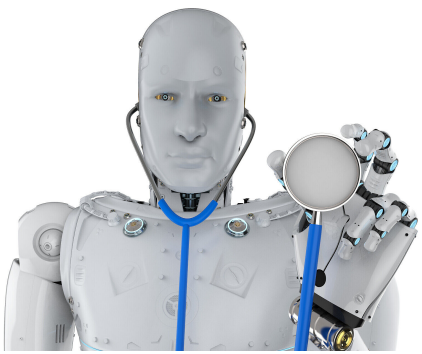
Harvard University

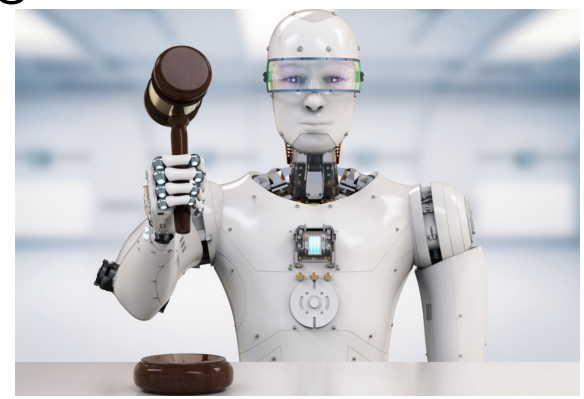
Japanese Society for Quantitative Political Science
Kochi University of Technology
July 6, 2024

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 algorithms are used throughout our society
- 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
  - How do AI recommendations influence human decisions?
  - Does AI help humans make more accurate decisions?

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
  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
- Field experiment
  - randomized provision of PSA to a judge 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] **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

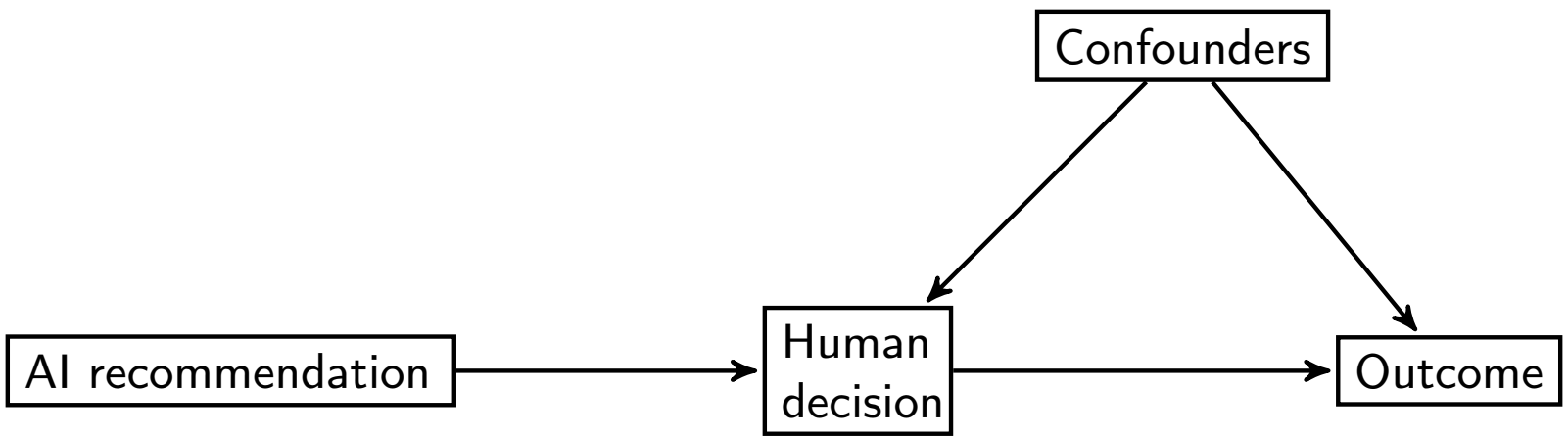
Does the Judge Agree with AI?

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

		AI	
		Signature bond	Cash bail
Human+AI	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 blindness: 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
- 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

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:

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

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

Comparing Human Decisions with and without AI

- Confusion matrix:

$$C(D(z)) = \begin{bmatrix} p_{00}(D(z)) & p_{01}(D(z)) \\ p_{10}(D(z)) & p_{11}(D(z)) \end{bmatrix}$$

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:

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

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

- Identification result:

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

- 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 Decisions

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

$$C(A) = \begin{bmatrix} p_{00}(A) & p_{01}(A) \\ p_{10}(A) & p_{11}(A) \end{bmatrix}$$

- Derive sharp bounds of the risk differences:

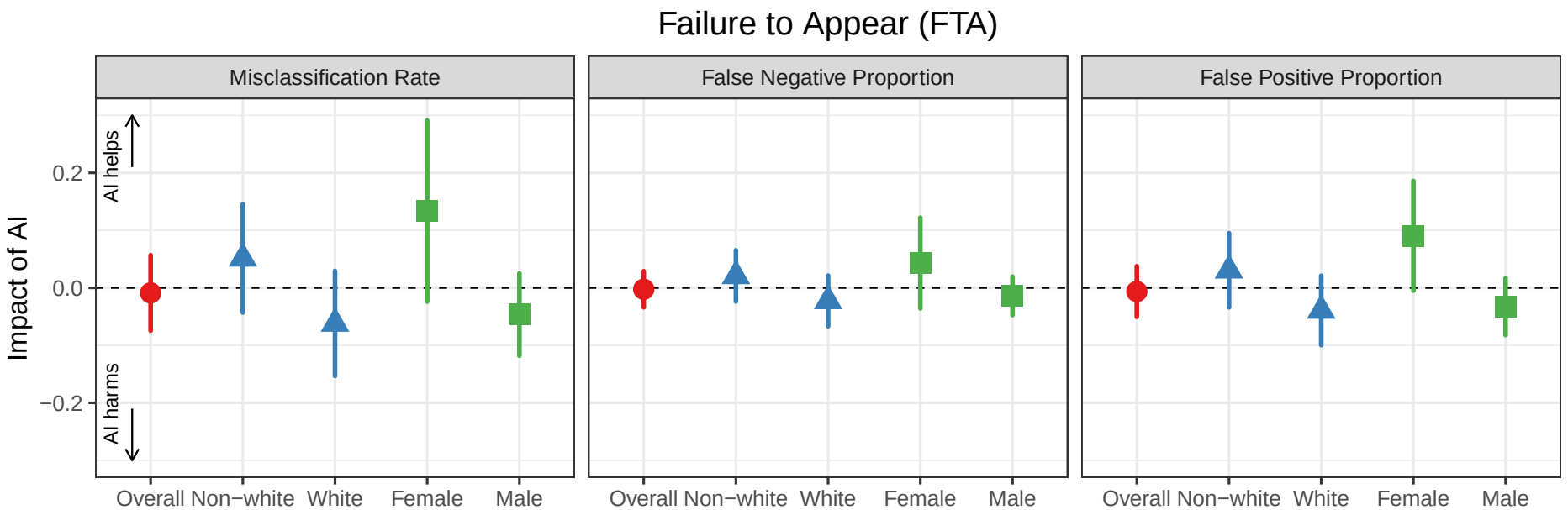
$$R_{\text{AI}}(\ell_{01}) - R_{\text{Human}}(\ell_{01}) \quad \text{and} \quad R_{\text{AI}}(\ell_{01}) - R_{\text{Human+AI}}(\ell_{01}),$$

using

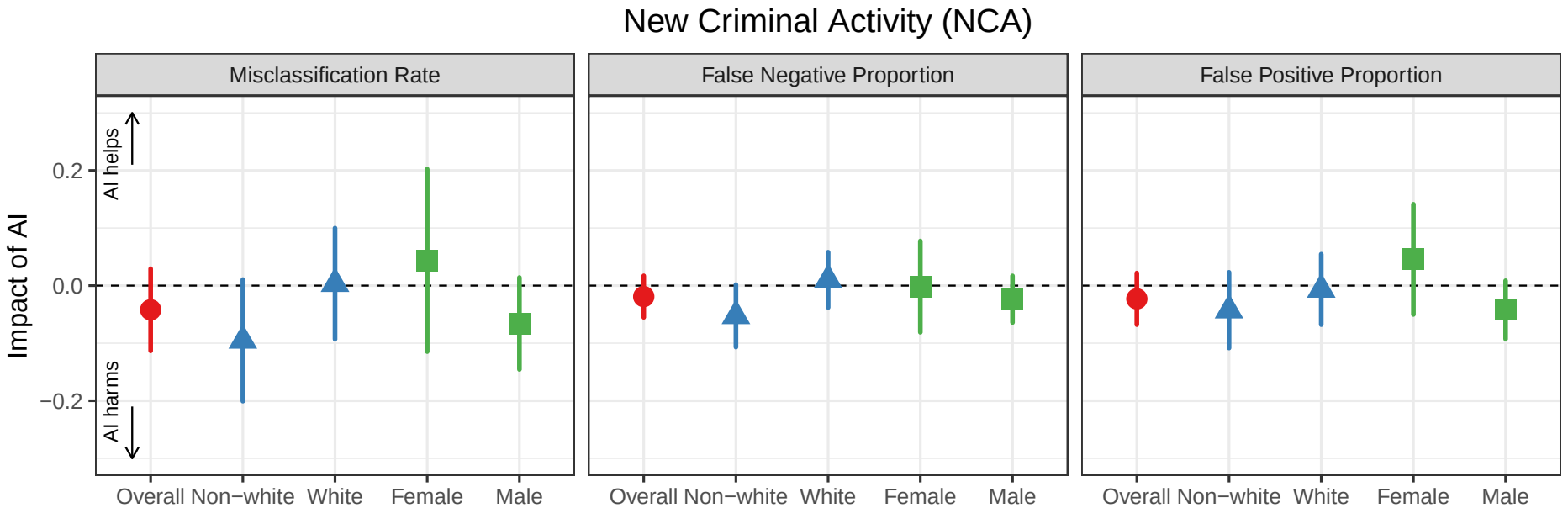
$$p_{ya}(A) = \Pr(Y(0) = y, D = 1, A = a) + \Pr(Y(0) = y, D = 0, A = a)$$

- Extend these methods to observational studies (double machine learning) under unconfoundedness $\{Y(d), D(z)\}_{d,z\in\{0,1\}} \perp\!\!\!\perp Z \mid X$

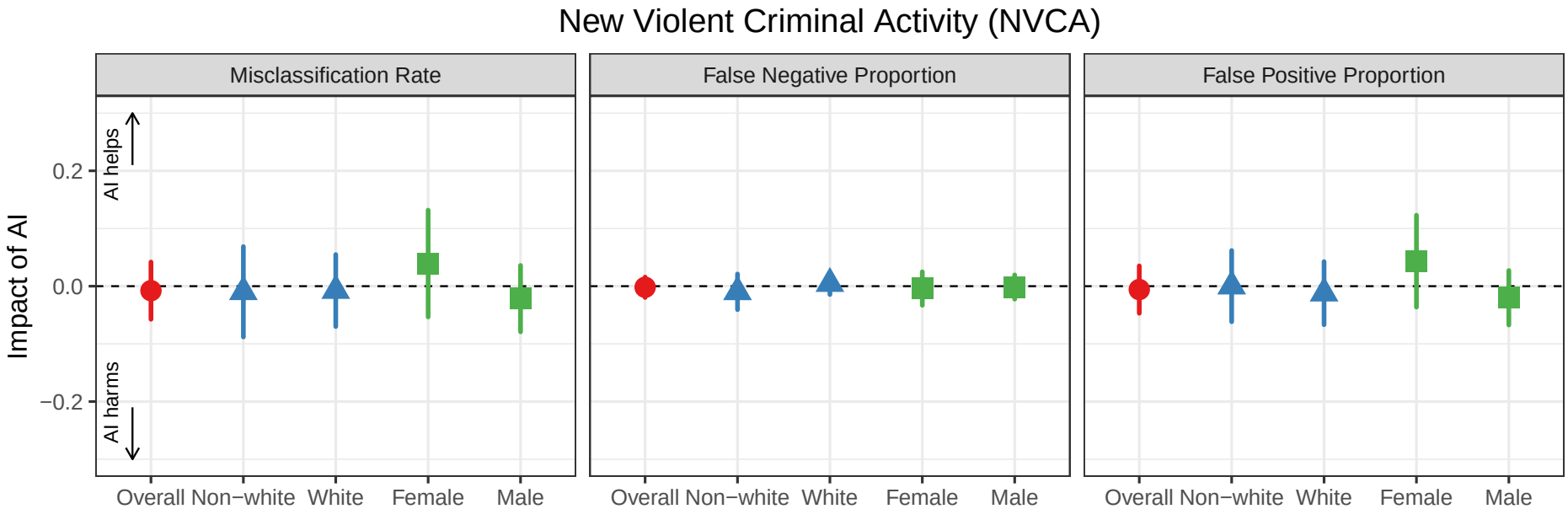
AI Recommendations Do Not Improve Human Decisions



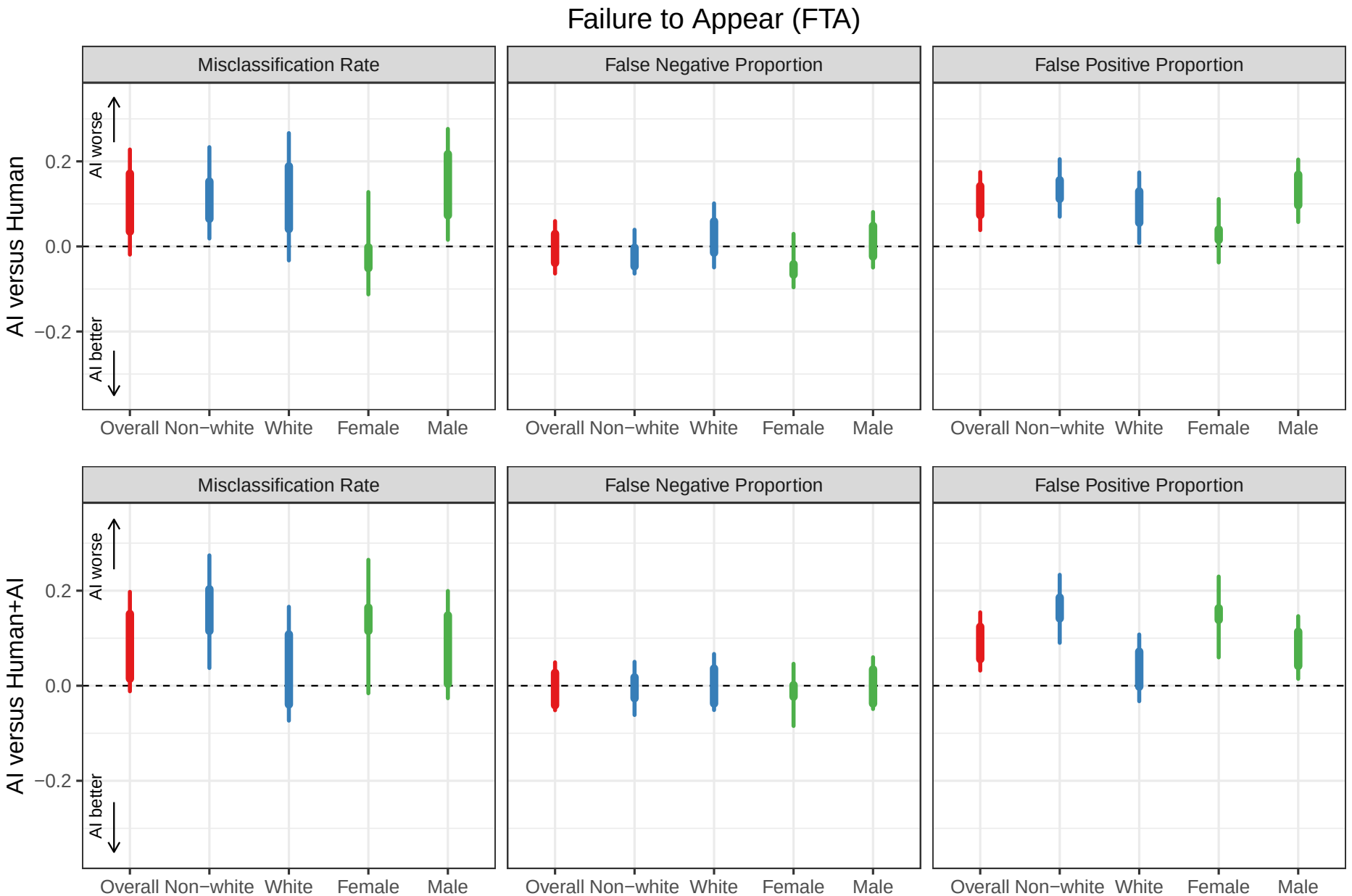
AI Recommendations Do Not Improve Human Decisions



AI Recommendations Do Not Improve Human Decisions

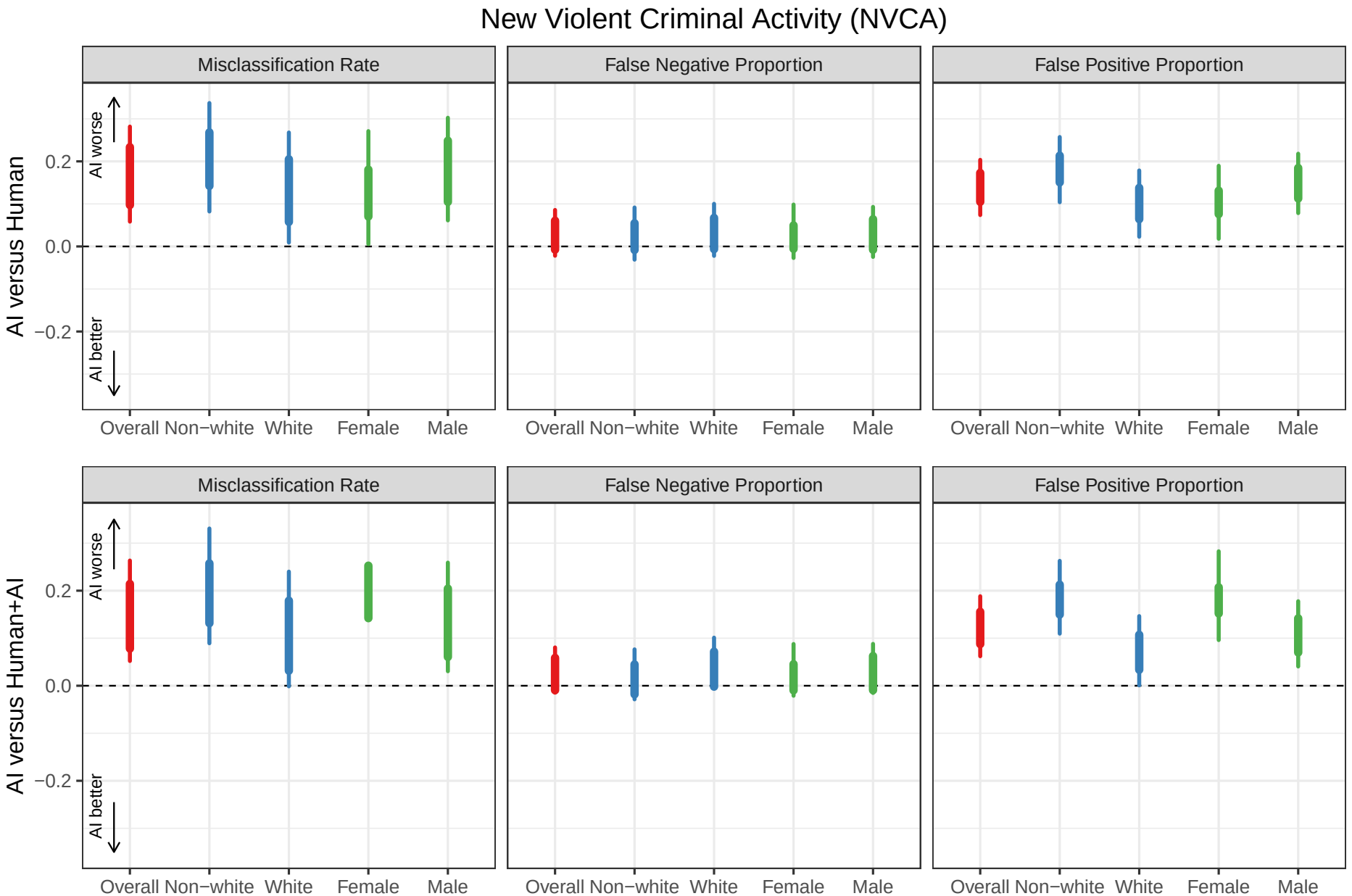


AI-Alone Decisions Perform Worse than Human Decisions

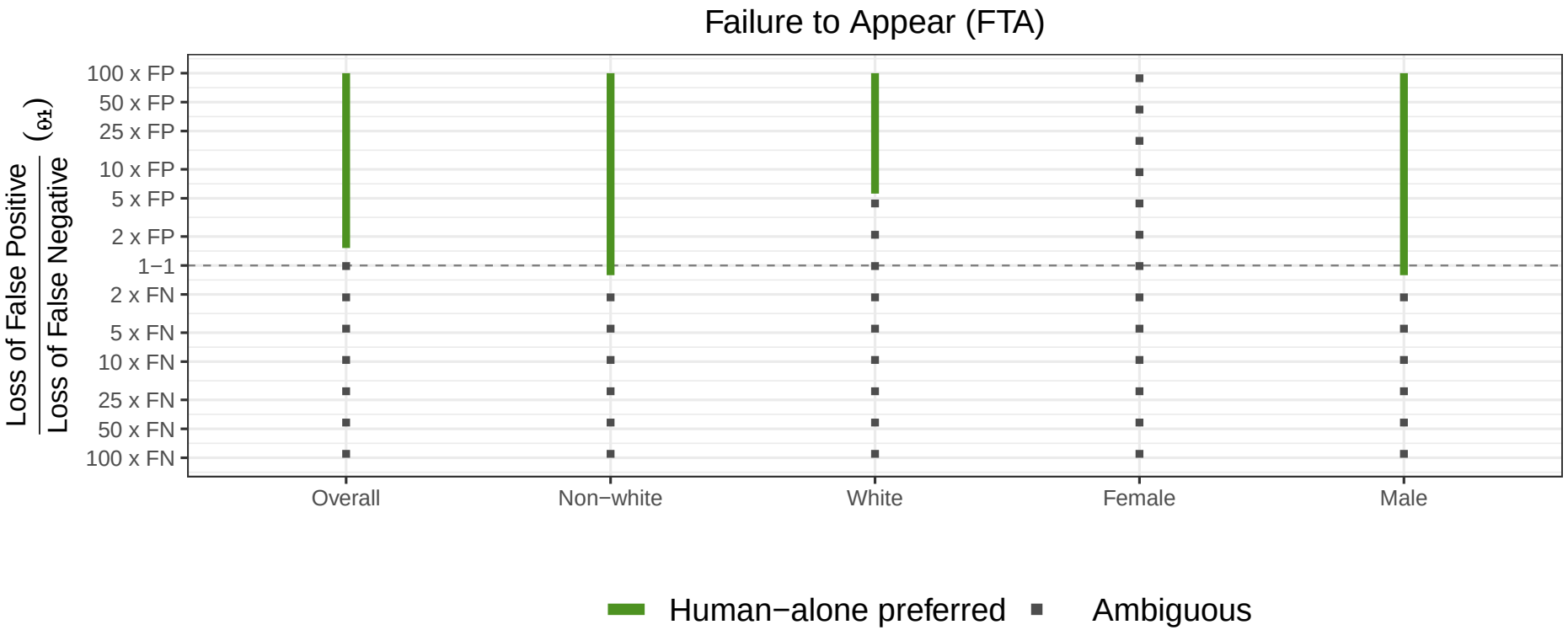


AI-Alone Decisions Perform Worse than Human Decisions

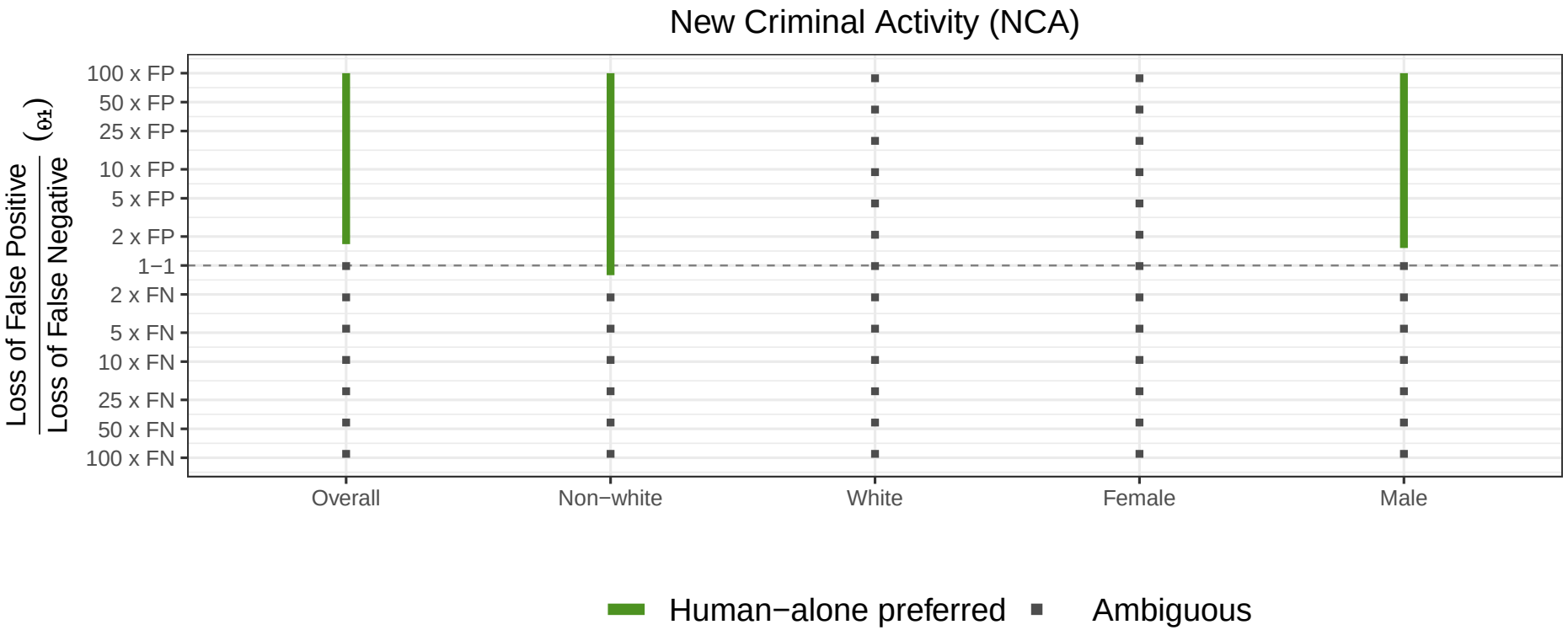
AI-Alone Decisions Perform Worse than Human Decisions



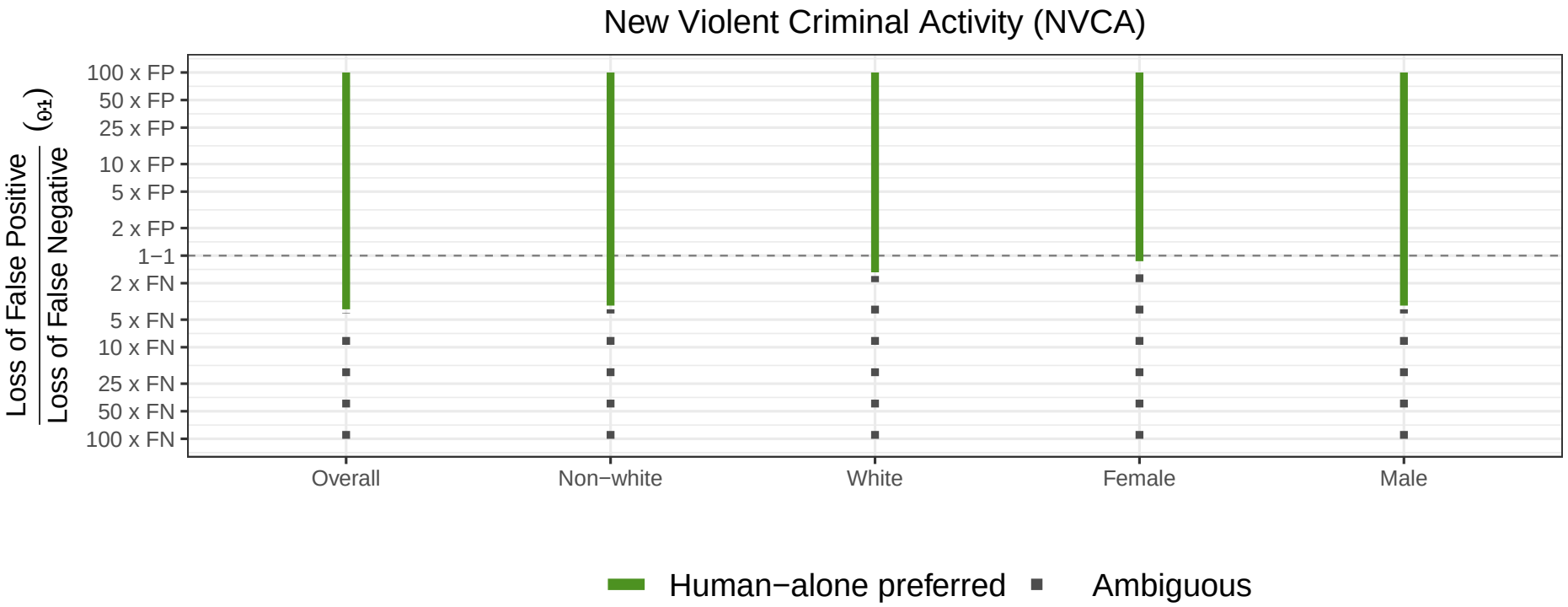
Human-Alone System is Preferred over AI-Alone System when the Cost of False Positive is High



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



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