Essential Role of Causality in the Fairness Evaluation of AI-Assisted Human Decision Making

Kosuke Imai

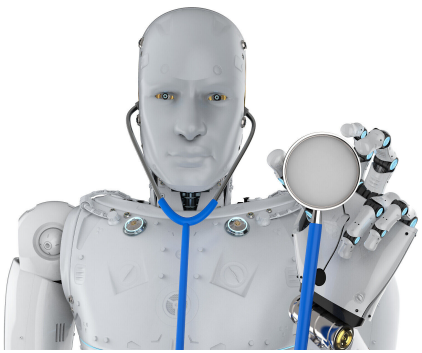
Harvard University

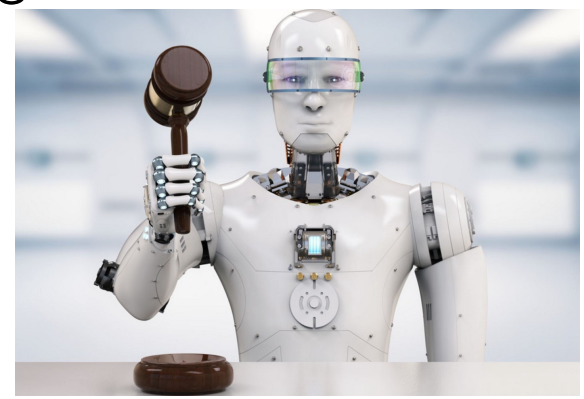
Symposium on Causality
Florence, Italy
September 27, 2024

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

- 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

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: classifies arrestees according to FTA and NCA/NVCA risks

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
  - 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: randomization of PSA provision

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

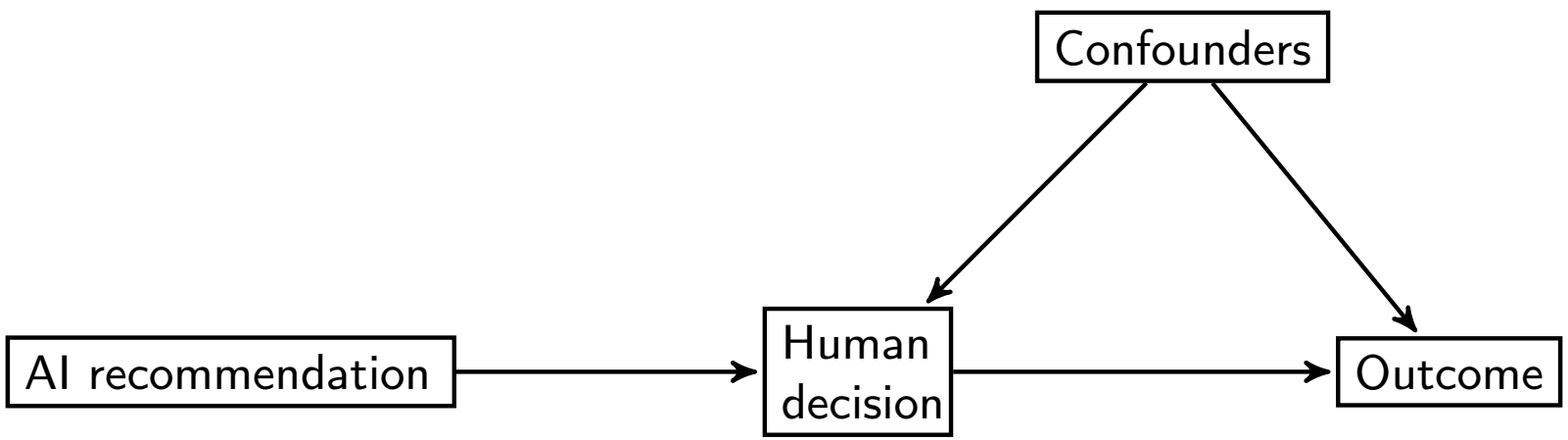
Does the Judge Agree with AI?

		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 blind: AI recommendations affect the outcome only through human decisions
- The proposed design is widely applicable even when stakes are high



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

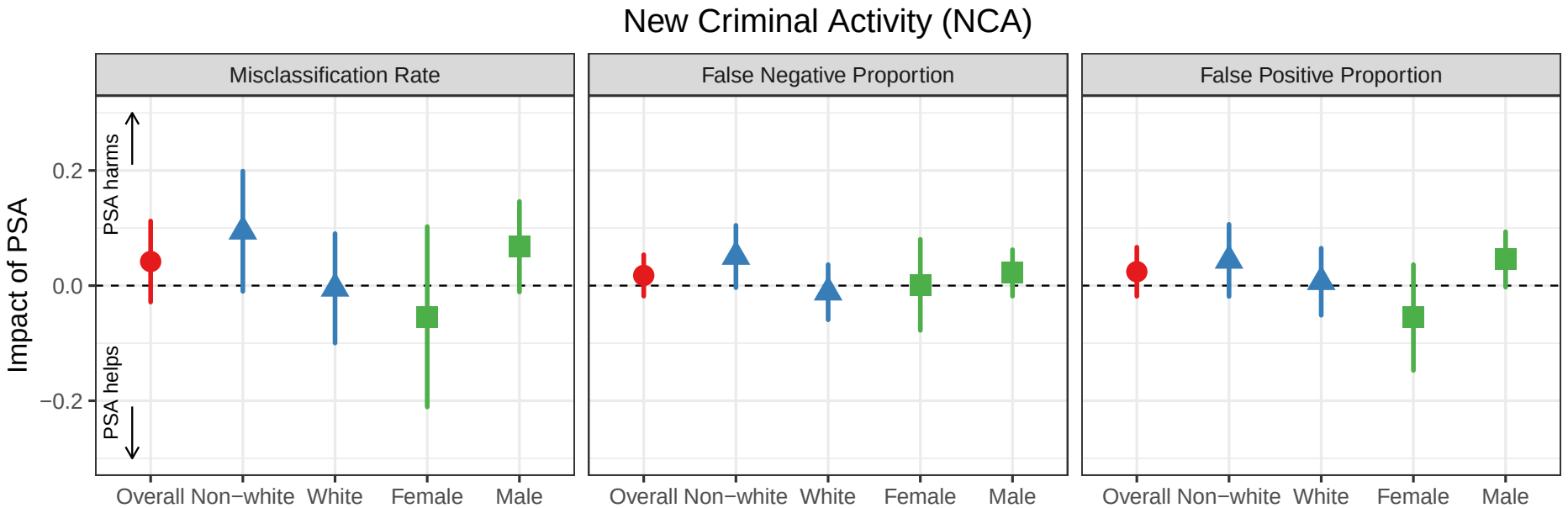
		Decision	
		Negative ($D = 0$)	Positive ($D = 1$)
Outcome	Negative ($Y(0) = 0$)	True Negative (TN) ℓ_{00}	False Positive (FP) ℓ_{01}
	Positive ($Y(0) = 1$)	False Negative (FN) $\ell_{10} = 1$	True Positive (TP) ℓ_{11}

- Assign a 'loss' to each classification outcome
- Classification risk:

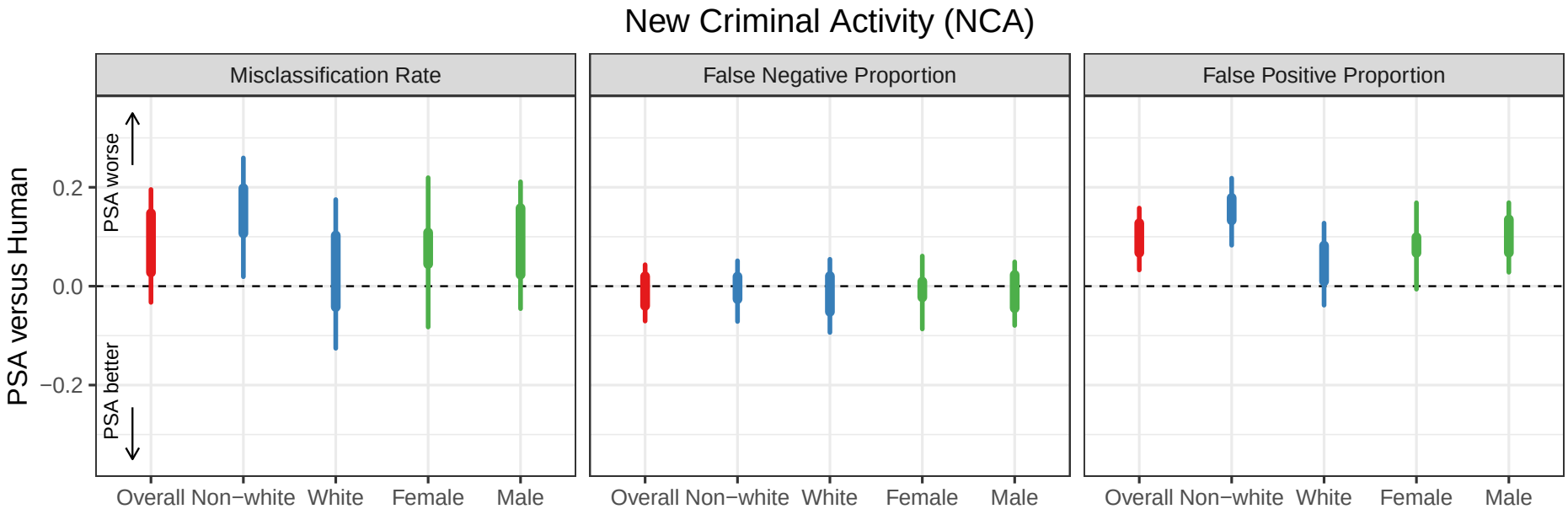
$$R(\ell_{01}) = \underbrace{\ell_{10}}_{=1} \cdot \text{FNP} + \ell_{01} \cdot \text{FPP}.$$

  where misclassification rate is $R(1) = \text{FNP} + \text{FPP}$
- We can identify the risk *difference* between Human vs. Human+AI
- We can bound the risk difference between Human vs. AI-alone

PSA Recommendations Do Not Improve Human Decisions



PSA-Alone Decisions Perform Worse than Human Decisions



Concluding Remarks

- Humans (still) make most high-stakes decisions
  - need to examine how AI affects human decisions
  - accurate/fair AI does not imply accurate/fair human decisions
- Causality plays an essential role
  - AI recommendations affect human decisions
  - human decisions influence outcomes

- We propose a methodological framework for experimentally evaluating the three decision-making systems:
  1. human-alone
  2. human+AI
  3. AI-alone
- We conducted and analyzed an RCT that evaluates the pretrial risk assessment instrument (PSA-DMF sytem):
  1. PSA recommendations have little impacts on human decisions
  2. PSA decisions perform worse than human decisions