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

Kosuke Imai
Harvard University
Quantitative Collaborative
University of Virginia
March 13, 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
Al-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
  3. First ever field experiment: evaluating pretrial public safety assessment
Controversy over the COMPAS Score (Propublica)

Two Petty Theft Arrests

VERNON PRATER
LOW RISK 3

BRISHA BORDEN
HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid’s bike and scooter that were sitting outside. She did not reoffend.

Two Drug Possession Arrests

DYLAN FUGETT
LOW RISK 3

BERNARD PARKER
HIGH RISK 10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

White Defendants’ Risk Scores

<table>
<thead>
<tr>
<th>Risk Score</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Black Defendants’ Risk Scores

<table>
<thead>
<tr>
<th>Risk Score</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>
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
  - 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
<table>
<thead>
<tr>
<th>Risk Factors</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Age at Current Arrest</td>
<td>23 or Older</td>
</tr>
<tr>
<td>2. Current Violent Offense</td>
<td>No</td>
</tr>
<tr>
<td>a. Current Violent Offense &amp; 20 Years Old or Younger</td>
<td>No</td>
</tr>
<tr>
<td>3. Pending Charge at the Time of the Offense</td>
<td>No</td>
</tr>
<tr>
<td>4. Prior Misdemeanor Conviction</td>
<td>Yes</td>
</tr>
<tr>
<td>5. Prior Felony Conviction</td>
<td>Yes</td>
</tr>
<tr>
<td>a. Prior Conviction</td>
<td>Yes</td>
</tr>
<tr>
<td>6. Prior Violent Conviction</td>
<td>2</td>
</tr>
<tr>
<td>7. Prior Failure to Appear Pretrial in Past 2 Years</td>
<td>0</td>
</tr>
<tr>
<td>8. Prior Failure to Appear Pretrial Older than 2 Years</td>
<td>Yes</td>
</tr>
<tr>
<td>9. Prior Sentence to Incarceration</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Recommendations:**

Release Recommendation - Signature bond
Conditions - Report to and comply with pretrial supervision
## PSA Provision, Demographics, and Outcomes

<table>
<thead>
<tr>
<th></th>
<th>no PSA</th>
<th>PSA</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Signature bond</td>
<td>Cash bail</td>
<td>Signature bond</td>
<td>Cash bail</td>
<td>Total (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>small</td>
<td>large</td>
<td>small</td>
<td>large</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-white female</td>
<td>64</td>
<td>11</td>
<td>6</td>
<td>6</td>
<td>154 (8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White female</td>
<td>91</td>
<td>17</td>
<td>7</td>
<td>7</td>
<td>246 (13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-white male</td>
<td>261</td>
<td>56</td>
<td>49</td>
<td>49</td>
<td>734 (39)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White male</td>
<td>289</td>
<td>48</td>
<td>44</td>
<td>44</td>
<td>757 (40)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTA committed</td>
<td>218</td>
<td>42</td>
<td>16</td>
<td>16</td>
<td>558 (29)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not committed</td>
<td>487</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>1333 (71)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCA committed</td>
<td>211</td>
<td>39</td>
<td>14</td>
<td>14</td>
<td>523 (28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not committed</td>
<td>494</td>
<td>93</td>
<td>92</td>
<td>92</td>
<td>1368 (72)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVCA committed</td>
<td>36</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>109 (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not committed</td>
<td>669</td>
<td>122</td>
<td>103</td>
<td>103</td>
<td>1782 (94)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (%)</td>
<td>705</td>
<td>132</td>
<td>106</td>
<td>106</td>
<td>1891 (100)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

*Note:* The numbers in parentheses represent the percentage of the total.
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?

<table>
<thead>
<tr>
<th></th>
<th>AI</th>
<th>Human</th>
<th>Human+AI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Signature bond</td>
<td>Cash bond</td>
<td>Signature bond</td>
</tr>
<tr>
<td>Signature bond</td>
<td>54.1% (510)</td>
<td>20.7 (195)</td>
<td>57.3% (543)</td>
</tr>
<tr>
<td>Cash bond</td>
<td>9.4 (89)</td>
<td>15.8 (149)</td>
<td>7.4 (70)</td>
</tr>
</tbody>
</table>
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:**
     \[
     Z_i \perp \{A_i, D_i(0), D_i(1), Y_i(0), Y_i(1)\} \quad \text{for all } i
     \]

- 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

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Positive ((D = 1))</td>
</tr>
<tr>
<td>Positive</td>
<td>Negative ((D = 0))</td>
</tr>
</tbody>
</table>

- True Negative (TN): No NCA
- False Positive (FP): Unnecessary cash bail
- False Negative (FN): Signature bond followed by NCA
- True Positive (TP): NCA

### Decision
- Positive: Cash bail
- Negative: Signature bond

### Outcome
- Positive: NCA
- Negative: No NCA
### Classification Risk

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive ($Y(0)=1$)</td>
<td>False Negative (FN)</td>
</tr>
<tr>
<td></td>
<td>True Positive (TP)</td>
</tr>
<tr>
<td>Negative ($Y(0)=0$)</td>
<td>True Negative (TN)</td>
</tr>
<tr>
<td></td>
<td>False Positive (FP)</td>
</tr>
</tbody>
</table>

- Assign a (possibly asymmetric) ‘loss’ to each classification outcome
- **Classification risk:**
  
  $$R(\ell_{01}) = \ell_{10} \cdot \text{FNP} + \ell_{01} \cdot \text{FPP} = q_{10} + \ell_{01} \cdot q_{01},$$
  
  where $q_{yd} = \Pr(Y(0) = y, D = d)$ for $y, d \in \{0, 1\}$

- **Other classification ability measures:**
  - misclassification rate: $R(1) = \text{FNP} + \text{FPP}$
  - $\text{FNR} = q_{10}/(q_{10} + q_{11})$, $\text{FPR} = q_{01}/(q_{00} + q_{01})$
  - false discovery rate: $\text{FDR} = q_{01}/(q_{01} + q_{11})$
Comparing Human Decisions with and without AI

- Define:
  \[ p_{yda}(z) := \Pr(Y(0) = y, D(z) = d, A = a) \]

- Confusion matrix:
  \[
  C_{\text{Human}}(z) = \begin{bmatrix}
  p_{000}(z) + p_{001}(z) & p_{010}(z) + p_{011}(z) \\
  p_{100}(z) + p_{101}(z) & p_{110}(z) + p_{111}(z)
  \end{bmatrix}
  \]
  \[
  = \begin{bmatrix}
  p_{00}(z) & p_{01}(z) \\
  p_{10}(z) & p_{11}(z)
  \end{bmatrix}
  \]
  marginalize over AI recommendations

  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:

\[
\Pr(Y(0) = 0 \mid Z = 1) = \Pr(Y(0) = 0 \mid Z = 0) \quad \text{by randomization}
\]

\[
\begin{align*}
p_{01}(1) + p_{00}(1) & = p_{01}(0) + p_{00}(0) \\
p_{01}(1) - p_{01}(0) & = p_{00}(0) - p_{00}(1)
\end{align*}
\]

- Identification result:

\[
R_{\text{Human+AI}}(\ell_{01}) - R_{\text{Human}}(\ell_{01})
\]

\[
= (p_{10}(1) + \ell_{01}p_{01}(1)) - (p_{10}(0) + \ell_{01}p_{01}(0))
\]

\[
= p_{10}(1) - p_{10}(0) + \ell_{01}(p_{00}(0) - p_{00}(1))
\]

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

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

- Invert this test to obtain a confidence interval on \(\ell_{01}\)
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

\[
C_{AI} = \begin{bmatrix}
p_{000}(z) + p_{010}(z) & p_{001}(z) + p_{011}(z) \\
p_{100}(z) + p_{110}(z) & p_{101}(z) + p_{111}(z)
\end{bmatrix}
= \begin{bmatrix}
p_{0\cdot0}(z) & p_{0\cdot1}(z) \\
p_{1\cdot0}(z) & p_{1\cdot1}(z)
\end{bmatrix}
\]

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

\[
p_{y1a}(z) = \Pr(Y(0) = y \mid D(z) = 1, Z = z, A = a) \\
\in [0,1] \\
\times P(D(z) = 1 \mid A = a, Z = z) \cdot \Pr(A = a) \\
\in [0, \Pr(D = 1 \mid A = a, Z = z) \Pr(A = a)]
\]
AI Recommendations Do Not Improve Human Decisions

Failure to Appear (FTA)

- **Misclassification Rate**
  - AI harms
  - AI helps

- **False Negative Proportion**

- **False Positive Proportion**

<table>
<thead>
<tr>
<th>Impact of AI</th>
<th>Overall Non-white</th>
<th>White</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI harms</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI helps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Overall Non-white</th>
<th>White</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misclassification Rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Negative Proportion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Positive Proportion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
AI Recommendations Do Not Improve Human Decisions

New Criminal Activity (NCA)

<table>
<thead>
<tr>
<th>Misclassification Rate</th>
<th>False Negative Proportion</th>
<th>False Positive Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>Non-white</td>
<td>White</td>
</tr>
<tr>
<td>Impact of AI</td>
<td>Impact of AI</td>
<td>Impact of AI</td>
</tr>
<tr>
<td>AI helps</td>
<td>AI harms</td>
<td>AI helps</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>
AI Recommendations Do Not Improve Human Decisions

- Misclassification Rate
- False Negative Proportion
- False Positive Proportion

Impact of AI

New Violent Criminal Activity (NVCA)
AI-Alone Decisions Perform Worse than Human Decisions

Failure to Appear (FTA)

- Misclassification Rate
- False Negative Proportion
- False Positive Proportion

Overall Non−white | White | Female | Male
---|---|---|---
AI worse | 0.2 | 0.0 | 0.2
AI better | -0.2 | 0.0 | 0.2

AI versus Human

AI versus Human+AI

Overall Non−white | White | Female | Male
---|---|---|---
AI worse | 0.2 | 0.0 | 0.2
AI better | -0.2 | 0.0 | 0.2
AI-Alone Decisions Perform Worse than Human Decisions

New Criminal Activity (NCA)

Misclassification Rate
False Negative Proportion
False Positive Proportion

Overall Non−white White Female Male Overall Non−white White Female Male Overall Non−white White Female Male

AI versus Human

AI versus Human+AI

AI worse
AI better

Overall Non−white White Female Male Overall Non−white White Female Male Overall Non−white White Female Male
AI-Alone Decisions Perform Worse than Human Decisions

New Violent Criminal Activity (NVCA)

<table>
<thead>
<tr>
<th></th>
<th>Misclassification Rate</th>
<th>False Negative Proportion</th>
<th>False Positive Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non−white</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI worse</td>
<td>0.2</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>AI better</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

AI versus Human

AI versus Human+AI
AI-Alone System Has More False Positives for Non-whites

### Failure to Appear (FTA)

<table>
<thead>
<tr>
<th></th>
<th>AI</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Misclassification Rate</strong></td>
<td><img src="image" alt="Graph AI Misclassification Rate" /></td>
<td><img src="image" alt="Graph Human Misclassification Rate" /></td>
</tr>
<tr>
<td><strong>False Negative Proportion</strong></td>
<td><img src="image" alt="Graph AI False Negative Proportion" /></td>
<td><img src="image" alt="Graph Human False Negative Proportion" /></td>
</tr>
<tr>
<td><strong>False Positive Proportion</strong></td>
<td><img src="image" alt="Graph AI False Positive Proportion" /></td>
<td><img src="image" alt="Graph Human False Positive Proportion" /></td>
</tr>
</tbody>
</table>

*Overall, Non-white, White, Female, Male*
AI-Alone System Has More False Positives for Non-whites

New Criminal Activity (NCA)

<table>
<thead>
<tr>
<th></th>
<th>Misclassification Rate</th>
<th>False Negative Proportion</th>
<th>False Positive Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall Non-white</td>
<td>White</td>
<td>Female</td>
</tr>
<tr>
<td>AI</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>Human</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

26 / 28
AI-Alone System Has More False Positives for Non-whites

New Violent Criminal Activity (NVCA)

Misclassification Rate

False Negative Proportion

False Positive Proportion

Overall Non−white White Female Male Overall Non−white White Female Male
0.0
0.1
0.2
0.3
0.4
0.5AI
New Violent Criminal Activity (NVCA)

Misclassification Rate

False Negative Proportion

False Positive Proportion

Overall Non−white White Female Male Overall Non−white White Female Male
0.0
0.1
0.2
0.3
0.4
0.5Human
New Violent Criminal Activity (NVCA)

Misclassification Rate

False Negative Proportion

False Positive Proportion

Overall Non−white White Female Male Overall Non−white White Female Male
0.0
0.1
0.2
0.3
0.4
0.5Human
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 system):

1. AI recommendations have little impacts on human decisions
2. AI decisions perform worse than human decisions
## PSA Scoring Rule

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>FTA</th>
<th>NCA</th>
<th>NVCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current violent offense</td>
<td>&gt; 20 years old</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>≤ 20 years old</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Pending charge at time of arrest</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Prior conviction</td>
<td>misdemeanor or felony</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>misdemeanor and felony</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Prior violent conviction</td>
<td>1 or 2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3 or more</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Prior sentence to incarceration</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior FTA in past 2 years</td>
<td>only 1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2 or more</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Prior FTA older than 2 years</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>22 years or younger</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

- FTA: \( \{0 \rightarrow 1, 1 \rightarrow 2, 2 \rightarrow 3, (3, 4) \rightarrow 4, (5, 6) \rightarrow 5, 7 \rightarrow 6\} \)
- NCA: \( \{0 \rightarrow 1, (1, 2) \rightarrow 2, (3, 4) \rightarrow 3, (5, 6) \rightarrow 4, (7, 8) \rightarrow 5, (9, 10, 11, 12, 13) \rightarrow 6\} \)
- NVCA: \( \{(0, 1, 2, 3) \rightarrow 0, (4, 5, 6, 7) \rightarrow 1\} \)
Decision Making Framework (DMF)

[Graph showing the relationship between NCA Score and FTA Score, with PSA Recommendation for Signature Bond and Cash Bail indicated by different colors.]