PNAS RESEARCH ARTICLE | SOCIAL SCIENCES STATISTICS

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

Eli Ben-Michael[a], D. James Greiner[b], Melody Huang[c,d], Kosuke Imai[e,f,1], Zhichao Jiang[g], and Sooahn Shin[e]

Affiliations are included on p. 10.

Edited by Bin Yu, University of California, Berkeley, CA; received March 5, 2025; accepted August 13, 2025

The use of AI, or more generally data-driven algorithms, has become ubiquitous in today's society. Yet, in many cases and especially when stakes are high, humans still make final decisions. The critical question, therefore, is whether AI helps humans make better decisions compared to a human-alone or AI-alone system. We introduce a methodological framework to answer this question empirically with minimal assumptions. We measure a decision maker's ability to make correct decisions using standard classification metrics based on the baseline potential outcome. We consider a single-blinded and unconfounded treatment assignment, in which the provision of AI-generated recommendations is assumed to be randomized across cases, conditional on observed covariates, with final decisions made by humans. Under this study design, we show how to compare the performance of three alternative decision-making systems—human-alone, human-with-AI, and AI-alone. Importantly, the AI-alone system encompasses any individualized treatment assignment, including those not used in the original study. We also show when AI recommendations should be provided to a human-decision maker, and when one should follow such recommendations. We apply the proposed methodology to our own randomized controlled trial evaluating a pretrial risk assessment instrument. We find that the risk assessment recommendations do not improve the classification accuracy of a judge's decision to impose cash bail. Furthermore, replacing a human judge with algorithms—the risk assessment score and a large language model in particular—yields worse classification performance.

algorithmic decision making | policy learning | risk scores | recommendation systems | fairness

AI, or more broadly data-driven algorithms, have found a wide range of applications, including judicial decisions in the criminal justice system, treatment decisions in medicine, and recommendations in online advertising. And yet, in many settings and especially when stakes are high, humans still make final decisions. The critical question, therefore, is whether AI recommendations help humans make better decisions compared to a human alone or an AI alone (1).

Recent literature has largely focused on questions of whether AI recommendations themselves are accurate or biased (2–6). However, AI recommendations may not improve the accuracy of human decisions if, for example, the human decision-maker selectively ignores them (7–9). Similarly, the fairness of an AI-assisted human decision-making system depends on how the bias of the AI system interacts with that of the human decisions.

In this paper, we introduce a methodological framework for researchers to evaluate whether the provision of AI recommendations helps humans make better decisions. We formulate the notion of a decision-maker's "ability" as a classification problem under the potential outcomes framework of causal inference (10, 11). If AI recommendations are helpful, their provision should improve a human decision-maker's ability to correctly classify potential outcomes.

For example, when deciding whether to impose cash bail on an arrestee or to release them on their own recognizance (that is, released without depositing money with the court), a judge must balance public safety and efficient court administration against various costs of incarceration. Thus, the judge's decision-making ability can be defined as the degree to which they can correctly classify the as-yet unobserved arrestee's behavior upon release. A key question is, then, whether AI recommendations help a judge reduce classification error.

A primary methodological challenge in evaluating the impact of various decision-making systems is the selective labels problem; the decision-makers determine, by making endogenous decisions, which potential outcomes are observed (12). In the

Significance

In medicine, public policy, and other areas, human decision-makers are increasingly relying on AI-generated recommendations. To facilitate safe deployment of such algorithmic recommendation systems, we develop a statistical framework that rigorously evaluates the empirical performance of AI-assisted human decision-making systems. The proposed methodology enables comparison of the classification performance of human-alone, human-with-AI, and AI-alone decision-making systems. We also show how to optimally combine human decision makers and an AI-recommendation system. We apply our methodology to the first-ever randomized controlled trial assessing a recently developed pretrial risk assessment instrument designed to assist human judges in their decisions of whether to impose cash bail. Our framework can be used to evaluate a variety of AI-recommendation systems in both academia and industry.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: imai@harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2505106122/-/DCSupplemental.

Published September 17, 2025.

aforementioned example, because judges decide which arrestees to release on their own recognizance or subject to cash bail, we cannot observe whether an arrestee who received the cash bail decision would have committed a new crime if a judge were to have released them without additional conditions. The fact that human decisions may depend on factors unobservable to researchers makes this evaluation problem difficult and distinguishes it from the standard AB testing evaluation framework.

To overcome this selective labels problem, we consider an evaluation design where the provision of AI recommendations is assumed to be, possibly conditional on observed covariates, randomly assigned to human decision-makers across cases. Such an evaluation design is feasible even in settings where, for legal or other reasons, a human, rather than an AI system, must make the final decisions. We consider a single-blinded treatment assignment, which guarantees that the provision of AI recommendations affects the outcome only through human decisions.

Under this design, we show that, without additional assumptions, it is possible to point-identify the difference in classification risk, between the human-alone and human-with-AI systems, even though the risk of each decision-making system is unidentifiable. Moreover, although the proposed design does not include an AI-alone decision-making system as one of the treatment conditions, we derive sharp bounds on the classification ability differences between an AI-alone system and a human-alone or human+AI system. This approach enables us to evaluate the relative merit of any AI-alone system regardless of whether it was used in the study. We demonstrate that these bounds can be informative, allowing performance comparison between these decision-making systems without imposing additional assumptions.

Last, we derive optimal decision rules for when AI recommendations should be provided to a human decision-maker and when one should follow such recommendations. In the judicial example above, we may be interested in identifying the types of cases for which we should provide judges with AI recommendations. In addition, we may also study when judges should follow an AI recommendation and when they should ignore it.

Our empirical application is an experiment in which all identification assumptions are guaranteed by the design. However, the methodology is also applicable to observational studies under an additional assumption of unconfoundedness. Furthermore, the proposed methodological framework is useful for dealing with the selective labels problem in the statistical evaluation of any decision-making system. For example, our framework can be used to compare the ability of different human decision makers (13).

Finally, our methodology goes beyond the standard AB testing approach, which typically evaluates how different decision-making systems influence the outcome or the decision. Under the AB testing framework, it is difficult to directly relate the effects on the outcome with those on the decision, and to assess how these causal effects jointly characterize the quality of a decision. In contrast, the proposed framework uses the potential outcome to link these two causal effects through the concepts of false positives and false negatives (in our application, they correspond to unnecessary cash bail and own recognizance release followed by rearrest). Unlike the AB testing framework, therefore, our approach directly considers the relationship between decision and outcome (14).

1.1. Evaluation of the Public Risk Assessment Instrument. We apply the proposed methodology to a randomized controlled trial (RCT) to assess how an algorithmically generated pretrial risk assessment instrument, called the Public Safety Assessment (PSA), affects judges' decisions at a criminal first appearance hearing (1, 15). In Dane County, Wisconsin, where we conducted the RCT, a judge at a first appearance hearing must decide whether to release an arrestee on their own recognizance or to impose cash bail as a condition of release. In this county, own recognizance release is called a "signature bond"—a term we will use in this paper.

If a judge assigns an arrestee a signature bond, the arrestee need not deposit money with the court to achieve pretrial release. For cash bail, the arrested individual must deposit the specified amount with the court to be released. The decision between signature bond and cash bail does not conclusively determine whether an arrestee will achieve pretrial release. For example, some arrestees assigned cash bail in fact pay it, and thus obtain their pretrial freedom, while others assigned signature bonds remain incarcerated because immigration or other criminal justice authorities request a "hold" from the relevant jail.

The PSA provides information to the judge for each decision regarding the arrested individual's risk of 1) failure to appear (FTA) at subsequent court dates, 2) new criminal activity (NCA), and 3) new violent criminal activity (NVCA). In the RCT, the judge receives the PSA for a randomly selected subset of all first appearance/bail hearings (see ref. 15, for details of the PSA instrument and experiment).

The PSA provides three numerical scores that correspond to its classification of FTA, NCA, and NVCA risks. The FTA and NCA risk scores have a total of six levels, while the NVCA risk score is binary. Nine factors about prior criminal history as well as age, which is the only demographic factor, serve as inputs to construct the PSA's scores; neither race nor gender is an input. Finally, a deterministic formula called the decision-making framework (DMF) combines the PSA's three risk scores with other information, such as a jurisdiction's resources, to produce an overall recommendation of either cash bail or signature bond. This overall PSA-DMF recommendation is the focus of our analysis in this paper. We analyze the interim data from this experiment.*

Table 1 compares the PSA recommendations with the judge's decisions (left table) and with the human-with-PSA decisions (right). Each cell presents the proportion of the corresponding cases with the number of such cases in parentheses. We find that human decisions, with or without the PSA recommendations, do not always agree with the PSA recommendations. Indeed, the judge goes against the PSA recommendations in slightly more than 30% of the cases. In cases of disagreement, the PSA recommendations tend to be harsher than human decisions. When provided with PSA recommendations, the judge agrees with them more often (by approximately 5.6 percentage points with the SE of 2.0; see *SI Appendix*, Fig. S1). Even in these cases, there exists a substantial amount of disagreement between the human decisions and PSA recommendations. Similar to the human-alone vs. PSA-alone comparison, when they disagree, the PSA recommendations tend to be harsher than the human-with-PSA decisions.

We evaluate whether PSA recommendations improve judges' classification ability. In addition, we are interested in comparing the classification ability of the PSA-alone decisions with that of the human-alone decisions. As mentioned above, the key challenge is the presence of selective labels: For cases where the judge issued a cash bail decision, we do not observe the counterfactual outcome (FTA, NCA, NVCA) under a signature bond decision. The evaluation of the PSA-alone decision-making system in itself is even more difficult because the experiment does not have a PSA-alone condition. Our proposed methodological framework shows how to overcome these challenges without

*This is a slightly updated version of the data originally analyzed by ref. 1 and has been subsequently made publicly available by ref. 16.

Table 1. Comparison between human decisions and PSA-generated recommendations

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

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

The left table compares the PSA recommendations (columns) against the judge's decisions without PSA recommendations (rows). Similarly, the right table compares the PSA recommendations (columns) with the decisions made by a human judge who was provided with PSA recommendations (rows). Each cell presents the proportion of corresponding cases with the number of such cases in parentheses.

additional assumptions beyond those guaranteed by the experimental design.

Our empirical analysis shows that PSA recommendations do not significantly improve the classification ability of judge's decision-making. For example, the misclassification rate of judge's decisions is unchanged when the PSA recommendations are provided. We also find that PSA-alone decisions perform worse than human decisions. In particular, the PSA system yields a greater proportion of false positives (that is, imposing cash bail on an arrestee who would not commit a crime if released on their own recognizance).

1.2. Related Literature. The existing literature on algorithmic decision-making has primarily focused on three areas: i) the performance evaluation of algorithms in terms of their underlying classification tasks (e.g., refs. 17–20), ii) issues of algorithmic fairness and the potential for biased algorithmic or human recommendations (e.g., refs. 2–6, 21, and 22), and iii) understanding how humans incorporate algorithmic recommendations into their decision-making (e.g., refs. 1, 8, 9, 23, and 24).

We focus on a question at the intersection of these three areas. While other scholars have proposed using a classification framework with selective labels to consider the performance of algorithmic decision-making (e.g., refs. 18–20), we focus on the relative gains of AI recommendations over human decisions and show that they can be credibly identified under an RCT, even though each decision-making system cannot be evaluated in isolation. *SI Appendix*, Section S1 further explains our contributions to the related literature.

2. Materials and Methods

In this section, we introduce a methodological framework for evaluating statistically the relative performance of human-alone, human-with-AI, and AI-alone decision-making systems.

2.1. Design and Assumptions. Let A_i represent the binary AI-generated recommendation for case i. We assume that AI recommendations can be computed for all cases. In our application, $A_i = 1$ means that AI recommends cash bail while $A_i = 0$ indicates that AI recommends a signature bond. We use Z_i to denote the binary treatment variable, representing the provision of such an AI recommendation. In our experiment, $Z_i = 1$ indicates that a human judge receives a PSA recommendation, whereas $Z_i = 0$ means that no PSA recommendation is given to the judge. We will also assume that we observe case-level covariate information, denoted as $X_i \in \mathcal{X}$, where $\mathcal{X}$ is the support.

The proposed methodology can be generalized to settings with more than two treatment conditions. For example, researchers may use different AI recommendation systems or include an AI-alone decision system as a separate treatment arm.

We use D_i to denote the observed binary decision made by a human. Thus, $D_i = 1$ represents a judge's decision to require cash bail as opposed to a signature bond. Because the AI recommendation can affect the human decision, we use potential outcomes notation and denote the decision under the treatment condition $Z_i = z$ as $D_i(z)$. That is, $D_i(1)$ and $D_i(0)$ represent the decisions made with and without an AI recommendation, respectively. The observed decision, therefore, is given by $D_i = D_i(Z_i)$.

Last, let Y_i denote the binary outcome of interest. Without loss of generality, we assume that $Y_i = 1$ represents an undesired outcome relative to $Y_i = 0$. In our empirical application, this variable represents whether or not an arrestee FTA in court, or engages in NCA or NVCA.

We consider the use of single-blinded treatment assignment. In our application, single-blinding means that an arrestee does not know whether a judge receives an AI recommendation. In other words, we assume that the provision of an AI recommendation, or lack thereof, can affect the outcome only through the human decision. The assumption is violated if a judge informs an arrestee about the AI recommendation, which in turn affects the arrestee's behavior directly other than through the judge's decision. Formally, let $Y_i(z, d)$ denote the potential outcome under the treatment condition $Z_i = z$ and the decision $D_i = d$. The single-blinded experiment assumption implies $Y_i(0, d) = Y_i(1, d) = Y_i(d)$ for any d where the observed outcome is given by $Y_i = Y_i(D_i(Z_i))$.

In sum, our evaluation design requires unconfoundedness, overlap, and single-blinded treatment assignment.

Assumption 1. *The study design satisfies:*

1. *Single-blinded treatment assignment:* $Y_i(z, D_i(z)) = Y_i(z', D_i(z'))$ *for all* z, z' *such that* $D_i(z) = D_i(z')$
2. *Unconfounded treatment assignment:* $Z_i \perp\!\!\!\perp \{A_i, \{D_i(z), Y_i(d)\}_{z,d\in\{0,1\}}\} \mid X_i$
3. *Overlap:* $e(x) := \Pr(Z_i = 1 \mid X_i = x) \in [\eta, 1 - \eta]$ *for* $\eta > 0$

In RCTs, both unconfoundedness and overlap conditions are guaranteed to be satisfied. In observational studies, these conditions are assumed to hold. In addition, the treatment probabilities $e(x)$ are unknown and therefore must be estimated. We emphasize that the proposed methodology does not assume unconfoundedness and overlap regarding the human decision D (i.e., for all d, z, x, $Y_i(d) \perp\!\!\!\perp D_i \mid Z_i = z, X_i = x$ and $\Pr(D_i = d \mid Z_i = z, X_i = x) \in [\xi, 1 - \xi]$, where $\xi > 0$). This is an important advantage as human decisions often depend on factors that are unobservable to researchers.

The notation above implicitly assumes no spillover effects across cases. In our application, this means that a judge's decision should not be influenced by the treatment assignments of prior cases. To increase the credibility of this assumption, we focus on first arrest cases (see section S3 of *SI Appendix* of ref. 1 for empirical evidence consistent with this assumption). Finally, we assume that we have an independently identically distributed sample of cases with size n from a target distribution $\mathcal{P}$. In subsequent sections, we will omit the subscript i from expressions whenever convenient.

2.2. Measures of Classification Ability. We now formalize the "classification ability" of a decision-maker. We focus on the baseline potential outcome $Y(0)$. In our application, this corresponds to the outcome we would observe (e.g., NCA) if an arrestee is released on their own recognizance ($D = 0$). The fact that the PSA is designed to predict the behavior of an arrestee if released on their own recognizance also justifies the focus on $Y(0)$ in our application.

Table 2. Confusion matrix for each combination of baseline potential outcome $Y(0)$ and decision D^*

		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}

Each cell is assigned a loss ℓ_{yd} for $y, d \in \{0, 1\}$. The loss is standardized by setting $\ell_{10} = 1$.

Table 2 shows the confusion matrix for all four possible pairs of $Y(0)$ and a generic decision D^*, which is different from the observed human decision D in the study. If the baseline potential outcome is negative in the classification sense, i.e., $Y(0) = 0$ (e.g., an arrestee would not be rearrested for a new crime under a signature bond decision), and the decision is also negative $D^* = 0$ (e.g., a judge decides to assign a signature bond), then we call this instance a "true negative" or TN. In contrast, if the baseline outcome is positive (i.e., an undesired outcome in our application) and yet the decision is negative (e.g., a judge decides to release the arrestee on their own recognizance), then this instance is called "false negative" or FN (e.g., a person given a signature bond decision is rearrested for a new crime). False positives (FP; a person given a cash bail decision would not have been rearrested for a new crime under a signature bond decision) and true positives (TP; a person given a cash bail decision would be rearrested for a new crime under a signature bond decision) are similarly defined.

Using this confusion matrix, we can derive a range of classification ability measures. To do so, we first assign a loss (or negative utility) to each cell of the confusion matrix and then aggregate across cases. As shown in Table 2, let ℓ_{yd} denote a loss that is incurred when the baseline potential outcome is $Y(0) = y$ and the decision is $D^* = d$ for $y, d \in \{0, 1\}$. Without loss of generality (no pun intended), we set the loss of a false negative to one, i.e., $\ell_{10} = 1$.

This setup allows for both symmetric and asymmetric loss functions (25). If a false positive (e.g., unnecessary cash bail) and a false negative (e.g., signature bond, resulting in NCA) incur the same loss, i.e., $\ell_{01} = 1$, then the loss function is said to be symmetric. An asymmetric loss function arises, for example, if avoiding false negatives is deemed more valuable than preventing false positives, i.e., $\ell_{01} < 1$. To simplify the exposition, we will consider loss functions where true negatives and true positives incur zero loss (i.e., $\ell_{11} = \ell_{00} = 0$; see *SI Appendix*, Section S2 for a discussion of generic loss functions).

Once the loss function is defined, we compute the classification risk (or expected classification loss) as the average of the false negative proportion (FNP) and false positive proportion (FPP), weighted by their respective losses,

$$R(\ell_{01}; D^*) := p_{10}(D^*) + \ell_{01} p_{01}(D^*), \qquad [1]$$

where $p_{yd}(D^*) := \Pr(Y(0) = y, D^* = d)$ so that $p_{10}(D^*)$ and $p_{01}(D^*)$ represent the overall FNP and FPP, respectively, under a decision-making system D^*. When $\ell_{01} = 1$, the classification risk equals the misclassification rate, which represents the overall proportion of incorrect decisions.

We use this measure of classification ability to evaluate three decision-making systems: human-alone ($D^* = D(0)$), human-with-AI ($D^* = D(1)$), and AI-alone ($D^* = A$). We are particularly interested in contrasting the classification abilities of these three systems. For example, the comparison of human-alone and human-with-AI systems tells us whether AI recommendations are able to improve human decision-making.

One important limitation of our framework and related approaches, however, is that we only consider the baseline potential outcome rather than the joint potential outcomes. The "correct" or "wrong" decision might depend on both potential outcomes instead of the baseline potential outcome alone. Unfortunately, in general, the consideration of joint potential outcomes requires stronger assumptions than those considered under our approach (1). See *SI Appendix*, Section S1 and ref. 14 for a further discussion of related studies.

2.3. Comparing Human Decisions with and without AI Recommendations. We first show how to compare the performance of human decisions with and without AI recommendations under the above classification framework. We first derive the key identification result and then present our estimation strategy. We also propose a statistical hypothesis testing framework to compare different loss functions.

2.3.1. Identification. As explained in Section 2.2, our primary methodological challenge is the selective labels problem, which is commonly encountered in the evaluation of decisions. Specifically, we observe the baseline potential outcome under the negative decision $Y(0)$ only for cases where the decision is actually negative, i.e., $D = 0$.

Despite the selective labels problem, we show that it is possible to identify the difference in classification risk between human decisions with and without AI recommendations. To begin, the difference in classification risk between these two decision-making systems is given by,

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

where $R_{HUMAN}(\ell_{01}) := R(\ell_{01}; D(0))$ and $R_{HUMAN+AI}(\ell_{01}) := R(\ell_{01}; D(1))$ as defined in Eq. **1**. Under Assumption 1, we can immediately identify the effect of providing an AI recommendation on the FNP, i.e., $p_{10}(D(1)) - p_{10}(D(0))$.

Unfortunately, the FPP, $p_{01}(D(z))$, is not identifiable for $z = 0, 1$. Despite this fact, we can identify the average effect of access to AI recommendations on the FPP. Specifically, the sum of the FPP and the TNP equals $\Pr\{Y(0) = 0\}$ both with ($Z = 1$) and without ($Z = 0$) an AI recommendation:

$$\Pr\{Y(0) = 0\} = p_{01}(D(1)) + p_{00}(D(1)) = p_{01}(D(0)) + p_{00}(D(0)).$$

This implies that $p_{01}(D(1)) - p_{01}(D(0)) = p_{00}(D(1)) - p_{00}(D(0))$, which is identifiable under Assumption 1. The following theorem formally states the result.

Theorem 1. *Under Assumption 1, we can identify the difference in risk between human decisions with ($Z = 1$) and without ($Z = 0$) an AI recommendation as:*

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

2.3.2. Estimation. To estimate the difference in classification risk from the identification result in Theorem 1, we first write the identified form as the difference in means of a compound outcome: $W_i := Y_i(1 - D_i) - \ell_{01}(1 - Y_i)(1 - D_i)$. We can estimate this difference in classification risk via a variety of approaches, including the simple difference-in-means estimator.

Here, we consider a more general estimation approach based on the augmented inverse probability weighting (AIPW) estimator that can also be applied to observational studies (26). We begin by defining two nuisance components: i) the decision model $m^D(z, x) := \Pr(D = 1 \mid Z = z, X = x)$ and ii) the outcome model $m^Y(z, x) := \Pr(Y = 1 \mid D = 0, Z = z, X = x)$. For notational simplicity, we also define the propensity score under the treatment assignment z as $e(z, x) := ze(x) + (1 - z)\{1 - e(x)\}$ for a given pretreatment covariate value x, where, again, $e(x) := \Pr(Z = 1 \mid X = x)$. We assume that we estimate these nuisance components (and the propensity score $e(x)$) on a separate sample (Assumption S1 in *SI Appendix*, Section S4).

Once these nuisance components are estimated, we estimate the difference in classification risk as

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

where $\widehat{\varphi}_z$ are estimates of the (uncentered) influence function given by,

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

for $z = 0, 1$. We similarly define the true (uncentered) influence function as $\varphi_z(Z, X, D, Y; \ell_{01})$.

When rate and consistency conditions are satisfied, this estimator is asymptotically normally distributed around the true classification risk difference. Although for simplicity we assume that the nuisance components are fit on a separate sample, this is not necessary. All results readily extend to cross-fit estimators such as those we use in our application.

Theorem 2. *Under Assumption 1 and Assumption S1 of SI Appendix,*

$$\sqrt{n}\left[\hat{\beta} - \{R_{HUMAN+AI}(\ell_{01}) - R_{HUMAN}(\ell_{01})\}\right] \xrightarrow{d} N(0, V),$$

where $V = \mathbb{E}[\{\varphi_1(Z, X, D, Y; \ell_{01}) - \varphi_0(Z, X, D, Y; \ell_{01}) - (R_{HUMAN+AI}(\ell_{01}) - R_{HUMAN}(\ell_{01}))\}^2]$.

We can estimate the asymptotic variance as:

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\varphi}_1(Z_i, X_i, D_i, Y_i; \ell_{01}) - \widehat{\varphi}_0(Z_i, X_i, D_i, Y_i; \ell_{01}) - \hat{\beta}\right)^2.$$

Then, we obtain Wald-type $1-\alpha$ confidence intervals (CIs) as $\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\widehat{V}/n}$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution. The results of Theorem 2 can also be used to construct a power analysis (see *SI Appendix*, Section S14).

2.3.3. Comparing different loss functions. Whether one prefers the human decision-making system with or without AI recommendations depends on the chosen loss function (i.e., the value of ℓ_{01} in Eq. **1**). Using Theorem 1, we can ask under what loss functions we might prefer the human-with-AI decision-making system over the human-alone system.

We first consider the following hypothesis test that for a given ratio of the loss between false positives and false negatives ℓ_{01}, the risk is lower for the human-with-AI system:

$$\begin{aligned}
H_0 &: R_{HUMAN}(\ell_{01}) \leq R_{HUMAN+AI}(\ell_{01}), \\
H_1 &: R_{HUMAN}(\ell_{01}) > R_{HUMAN+AI}(\ell_{01}).
\end{aligned}$$

Inverting this hypothesis test for the parameter ℓ_{01} gives the values of the false positive loss for which we cannot rule out the possibility that the human-alone system is better than the human-with-AI system. Conversely, the region of ℓ_{01} where we reject H_0 gives the loss functions for which the human-alone system is unlikely to be better than the human-with-AI system.

Similarly, if we flip the null and alternative hypotheses so that H_1 becomes the null hypothesis, the region where we can reject it gives the relative values of the false positive loss that rule out the scenario that the human-with-AI system is better. The remaining cases are ambiguous.

2.4. Comparing a Generic AI Decision-Making System with Human Decisions. We next compare the classification ability of AI-alone decisions with human-alone and human-with-AI systems. The proposed approach is extremely general. In particular, we can analyze how any hypothetical AI-alone decision system would perform compared to a human-alone or human-with-AI system. This allows researchers to use data from a single study to evaluate different individualized decision rules, as long as the AI decision can be computed for any unit. For illustration, Section 3.6 compares the classification ability of a large language model with that of a human judge.

Unlike the comparison between human-alone and human-with-AI systems, we cannot point-identify the risk difference without imposing additional assumptions. This is because the proposed evaluation design does not have a treatment arm where an AI system makes decisions without human input. We do, however, observe AI recommendations for all cases since they can be readily computed. Below, we leverage this fact and derive informative bounds on the difference in classification risk between AI and human (with or without AI recommendations) decisions.

2.4.1. Partial identification. The fundamental problem here is that we do not observe the potential outcome under AI decisions $Y_i(A_i)$ when human decisions in the study (with or without AI recommendations) disagree with AI decisions, i.e., $A_i \neq D_i$. For example, Table 1 shows that in our application, the judge disagrees with AI recommendations in more than 25% of the cases. Furthermore, human decision-makers may disagree with AI decisions for reasons not observable by the researcher. To evaluate the AI-alone system, therefore, we must deal with this distinct selective labels problem.

The classification risk of the AI-alone system is defined as:

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

Because the study does not contain an AI-alone treatment arm, each term of the above equation is a mixture of identifiable and nonidentifiable parts:

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

where the first term is not identifiable. As a result, without further assumptions, the classification risk of an AI-alone system cannot be identified.

However, we can partially identify the differences in classification risk between AI-alone and human-alone/human-with-AI decision-making systems, focusing on the cases where AI recommendations differ from human decisions. Theorem 3 provides sharp (shortest possible) bounds on the range of possible values that risk differences can take on (see Theorems S1 and S2 in *SI Appendix*, Section S3 for the sharp bounds on the classification risk of a generic AI decision system).

Theorem 3. *Under Assumption 1, the risk differences are sharply bounded by the following:*

$$\begin{aligned}
\mathbb{E}[L_0(X)] \;&\leq\; R_{AI}(\ell_{01}) - R_{HUMAN}(\ell_{01}) \;&&\leq\; \mathbb{E}[U_0(X)], \\
\mathbb{E}[L_1(X)] \;&\leq\; R_{AI}(\ell_{01}) - R_{HUMAN+AI}(\ell_{01}) \;&&\leq\; \mathbb{E}[U_1(X)],
\end{aligned}$$

where the exact expressions of $L_z(x)$ *and* $U_z(x)$ *are given in SI Appendix, Section S5.*

The expressions of the lower and upper bounds involve the intersection of the bounds implied by each treatment arm, showing how randomization allows us to combine information across both arms. The width of the bounds is equal to,

$$\begin{aligned}
\mathbb{E}\{U_z(X) - L_z(X)\} = (1 + \ell_{01})\mathbb{E}\Big\{&\Pr(A = 0 \mid X) \\
&- \max_{z'}\Pr(Y = 1, D = 0, A = 0 \mid Z = z', X) \\
&- \max_{z'}\Pr(Y = 0, D = 0, A = 0 \mid Z = z', X)\Big\}.
\end{aligned}$$

The bounds tend to be narrow when the judge's decisions (with or without AI recommendations) align with the AI recommendations. Note that only cases with $D = 0$ matter because we focus on the potential outcome $Y(0)$.

2.4.2. Estimation. We now turn to estimation. Again, we consider a general approach that is applicable to observational studies. We define additional nuisance components corresponding to the decision and outcome models, while also conditioning on the AI recommendation A: $m^D(z,x,a) := \Pr(D = 1 \mid Z = z, X = x, A = a)$ and $m^Y(z,x,a) := \Pr(Y = 1 \mid D = 0, Z = z, X = x, A = a)$. Estimating the sharp bounds derived in Theorem 3 is complex, requiring i) the determination of which treatment choice z' achieves a tighter bound and ii) the estimation of the bound given the optimal choice of treatment arm z'.

Tackling the first component, for each covariate value $x \in \mathcal{X}$, we can characterize whether using $z' = z$ or $z' = 1 - z$ results in a greater lower bound with a nuisance classifier:

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

where $g_{L_z}(x) = 0$ denotes that the optimal choice is $z' = z$ and $g_{L_z}(x) = 1$ denotes that it is $z' = 1 - z$. Similarly, we can characterize the choice of z' for the least upper bound with another nuisance classifier,

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

We estimate these nuisance classifiers by first estimating $m^D(z,x,a)$ and $m^Y(z,x,a)$, then plugging these estimates into the formulas for the nuisance classifiers.

For the second step, we estimate the bound corresponding to the choice of z' using an efficient AIPW estimator. We do this by noting that we can write the conditional probabilities in Theorem 3 as conditional expectations of compound outcomes: $Y(1-D)(1-A)$, $(1-Y)(1-D)(1-A)$, $(1-A)D$, and $A(1-D)$. The final estimator uses the two sets of influence function estimates. We provide the exact expressions of these estimators of the sharp lower and upper bounds, denoted by $\widehat{L}_z$ and $\widehat{U}_z$, in *SI Appendix*, Section S6.

To estimate the bounds well, we need the plugin nuisance classifier to correctly classify which treatment arm to use for the bound. Systematic misclassification will lead to bias. One way to characterize the complexity of the classification problem is via a margin condition (27) that quantifies how often the difference between the two bounds is small. We formally state this margin condition as Assumption S2 in *SI Appendix*, Section S4. Together with a set of rate conditions presented as Assumption S3 in the same *SI Appendix*, we can establish the asymptotic normality of the estimated bounds.

Theorem 4. *Under Assumption 1 and Assumptions S1–S3 of SI Appendix, the estimated bounds are asymptotically normal,*

$$\sqrt{n}(\hat{L}_z - L_z) \xrightarrow{d} N(0, V_{L_z}), \quad \sqrt{n}(\hat{U}_z - U_z) \xrightarrow{d} N(0, V_{U_z}),$$

where the exact expressions of the asymptotic variances, V_{L_z} and V_{U_z}, are given in SI Appendix, Section S7.

Finally, to obtain CIs via Theorem 4, we first estimate the asymptotic variances by taking the required sample variances of estimated nuisance functions to obtain $\widehat{V}_{L_z}$ and $\widehat{V}_{U_z}$. We follow ref. 28 and compute the lower and upper $1-\alpha$ CIs for the lower and upper bounds, respectively. We then create a CI for the partially identified set as $\left[\widehat{L}_z - z_{1-\alpha}\sqrt{\widehat{V}_{L_z}/n},\ \widehat{U}_z + z_{1-\alpha}\sqrt{\widehat{V}_{U_z}/n}\right]$.

2.4.3. Comparing different loss functions. Similarly to Section 2.3, we can conduct a statistical hypothesis test to examine how the preference of the AI-alone system over the human-alone (or human-with-AI) system depends on the magnitude of loss ℓ_{01} assigned to false positives relative to false negatives. For example, to test whether the human-alone system is preferable to the AI-alone system, the null and alternative hypotheses are given by

$$H_0 : R_{AI}(\ell_{01}) \leq R_{HUMAN}(\ell_{01}), \quad H_1 : R_{AI}(\ell_{01}) > R_{HUMAN}(\ell_{01}). \quad [2]$$

If we reject H_0 for a given value of ℓ_{01}, the human-alone system is likely to have a lower risk than the AI-alone system.

Since the classification risk difference is only partially identified, we test the null hypothesis that its lower bound is less than or equal to zero, $H_0 : L_0 \leq 0$ vs. the alternative hypothesis $H_1 : L_0 > 0$. If we reject this null hypothesis, then we know that $R_{AI}(\ell_{01}) - R_{HUMAN}(\ell_{01}) \geq L_0 > 0$, implying that the risk of the AI-alone system is likely to be greater than that of the human-alone system and hence the latter is preferable. Similarly, if we reject the null hypothesis of $H_0 : U_0 \geq 0$ in favor of the alternative hypothesis $H_1 : U_0 < 0$, we prefer the AI-alone system over the human-alone system. As explained above, inverting these hypothesis tests will give us a range of loss functions under which the data either support preferring the human-alone or AI-alone systems (there will also be a region where the preference is ambiguous).

2.5. Policy Learning. We now consider whether these decision-making systems perform better in some cases than others, and how to derive rules for choosing which one to use. We first discuss learning when to provide AI recommendations and then analyze when human decision-makers should follow AI recommendations.

2.5.1. Learning when to provide AI recommendations. We first address the question of when to provide AI recommendations. The simplest approach would be to find a treatment policy that minimizes the expected number of negative outcome events, subject to a constraint on the maximum number of cash bail decisions. While this approach is reasonable, it is not clear how to specify this constraint in practice. Instead, we directly minimize the classification risk so that AI recommendations are provided only for the cases where they improve human decisions.

Let $\pi : \mathcal{X} \to \{0,1\}$ be a covariate-dependent policy that determines whether to provide the AI recommendation ($\pi(x) = 1$) or not ($\pi(x) = 0$). Here, the covariate space $\mathcal{X}$ may include the AI recommendation A. We consider a class of policies Π, each of which combines the human-alone and human-with-AI systems. The classification risk of policy $\pi \in \Pi$ is given by,

$$\begin{aligned} &R_{REC}(\ell_{01};\pi) \\ &:= p_{10}(D(\pi(X))) + \ell_{01}p_{01}(D(\pi(X))) \\ &= R_{HUMAN}(\ell_{01}) + \mathbb{E}\left[\pi(X)\left\{p_{10}(D(1) \mid X) - p_{10}(D(0) \mid X)\right.\right. \\ &\quad \left.\left. - \ell_{01} \times (p_{00}(D(1) \mid X) - p_{00}(D(0) \mid X))\right\}\right], \end{aligned}$$

where $p_{yd}(D^* \mid X) = \Pr(Y(0) = y, D^* = d \mid X)$ and we have used Theorem 1 to write the classification risk in terms of observable components.

Our goal is to find an optimal policy in Π that minimizes the classification risk,

$$\pi^*_{REC} \in \arg\min_{\pi\in\Pi} R_{REC}(\ell_{01};\pi). \quad [3]$$

We estimate this policy by solving the following empirical risk minimization problem with doubly robust estimators:

$$\hat{\pi}_{REC} \in \arg\min_{\pi\in\Pi} \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\left(\widehat{\varphi}_1(Z_i,X_i,D_i,Y_i;\ell_{01}) - \widehat{\varphi}_0(Z_i,X_i,D_i,Y_i;\ell_{01})\right).$$

The following theorem bounds the excess risk of this learned policy, i.e., the difference in classification risk between the best combined decision rule π^*_{REC} and the empirical rule $\hat{\pi}_{REC}$.

Theorem 5. *Under Assumption 1 and Assumptions S1–S3 of SI Appendix, we have*

$$\begin{aligned} &R_{REC}(\hat{\pi}_{REC};\ell_{01}) - R_{REC}(\pi^*_{REC};\ell_{01}) \\ &\leq C\left(\sum_{z=0}^{1}\|m^Y(z,\cdot) - \hat{m}^Y(z,\cdot)\|_2 + \|m^D(z,\cdot) - \hat{m}^D(z,\cdot)\|_2\right) \\ &\quad \times \|\hat{e} - e\|_2 + \left(1 + \frac{4}{\eta}\right)(1+\ell_{01})\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}}, \end{aligned}$$

with probability at least $1 - 2\exp(-t^2/2)$, where $\mathcal{R}_n(\Pi) := \mathbb{E}_{X,\epsilon}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\pi(X_i)\right|\right]$ is the population Rademacher complexity of the policy class Π.

Theorem 5 shows that the estimated policy will have low excess risk if the nuisance components are estimated well and the policy class is not too complex. The first component of the bound is related to the product-error rate seen in doubly robust policy learning (29). As in Section 2.3, due to the compound nature of the outcome, we can bound the excess risk in terms of the error rates for the outcome and decision models. In a randomized experiment, if we use the known propensity score (i.e., $\hat{e} = e$), this entire product term will disappear.

The analyst chooses the complexity of the policy class; flexible decision rules will be harder to estimate than simple ones, but simple, transparent rules are often preferred at the cost of potentially larger risk. As an example, if the policy class Π has a finite VC dimension v, the Rademacher complexity is $\mathcal{R}_n(\Pi) = O(\sqrt{v/n})$ (30, Section 5).

2.5.2. Learning when human decision-makers should follow AI recommendations. We next turn to learning when a human decision-maker should follow AI recommendations. We consider a policy π that determines when a human decision-maker should decide on their own ($\pi(x) = 0$) or simply follow the AI recommendation ($\pi(x) = 1$). The classification risk for a given policy π is a combination of the risk under the human-alone and the AI-alone system:

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

where $\widetilde{D} = A\pi(X) + D(0)(1 - \pi(X))$. As in Section 2.4, the expected risk has several unidentifiable terms. Therefore, we take a conservative approach and consider finding the decision rule that minimizes the worst-case excess risk relative to the human-alone system:

$$\pi^*_{DEC} \in \underset{\pi \in \Pi}{\arg\min}\ \mathbb{E}[\pi(X) U_0(X)], \qquad [4]$$

where $U_0(x)$ is the upper bound on the conditional risk difference derived in Theorem 3. This worst-case criterion requires strong evidence that the AI-alone decision is better before (hypothetically) overriding human decisions. It takes the human-alone decision as the baseline and only follows the AI recommendation if it will lead to a lower loss even in the worst case.

In *SI Appendix*, Section S8, we show how to minimize an estimate of the worst-case risk. We also analyze the error of this empirical risk minimization approach as done in Theorem 5.

3. Results

We now use the proposed methodology to analyze the experiment described in Section 1.1, focusing on evaluating three different decision-making systems—human-alone, PSA-alone, and human-with-PSA systems. Since we analyze only interim data, the results reported below should be interpreted as an illustration of the proposed methodology rather than the final analysis results from our RCT.

3.1. Setup. The dataset comprises a total of 1,891 first arrest cases, in which judges made decisions on whether to impose a signature bond ($D_i = 0$) or cash bail ($D_i = 1$). We dichotomize the PSA recommendation: $A_i = 1$ if it recommends a cash bail and $A_i = 0$ if the recommendation is a signature bond. We use the following case-level covariates X: gender (male or female), race (white or non-white), the interaction between gender and race, age, inputs for the PSA recommendation including variables for current/past charges and prior convictions, three PSA risk scores, and the overall PSA recommendation.

The provision of the PSA recommendation is randomized. In other words, the decision-maker in the treatment group is a human judge who is given the PSA recommendation ($Z_i = 1$), whereas in the control group, the same human judge makes decisions without the PSA recommendation ($Z_i = 0$). We use the true propensity score, $e(z, x) = 0.5$, for the estimation throughout the analysis.

Given the space constraint, we present the results for NCA, where $Y_i = 1$ indicates an incidence of NCA, and $Y_i = 0$ indicate absence. Among the cases, 40% are white males, 39% are non-white males, 13% are white females, and 8% are non-white females. The proportion of NCAs is 25%.

3.2. PSA Recommendations Do Not Improve Human Decisions. We begin by estimating the impact of providing PSA recommendations on human decisions. Specifically, we use the method described in Section 2.3 to estimate the difference in misclassification rates between decisions made by the human judge alone and those made with the PSA recommendation. Recall that the misclassification rate is equivalent to the symmetric loss function, i.e., $R(1; \mathrm{D}^*)$.

Fig. 1 presents the estimated impact of PSA recommendations on human decisions in terms of the misclassification rate, FNP, and FPP. We find that the PSA recommendations do not significantly improve the judge's decisions. Indeed, none of the classification risk differences between the judge's decisions with and without the PSA recommendations are statistically significant, though the estimates are relatively precise.

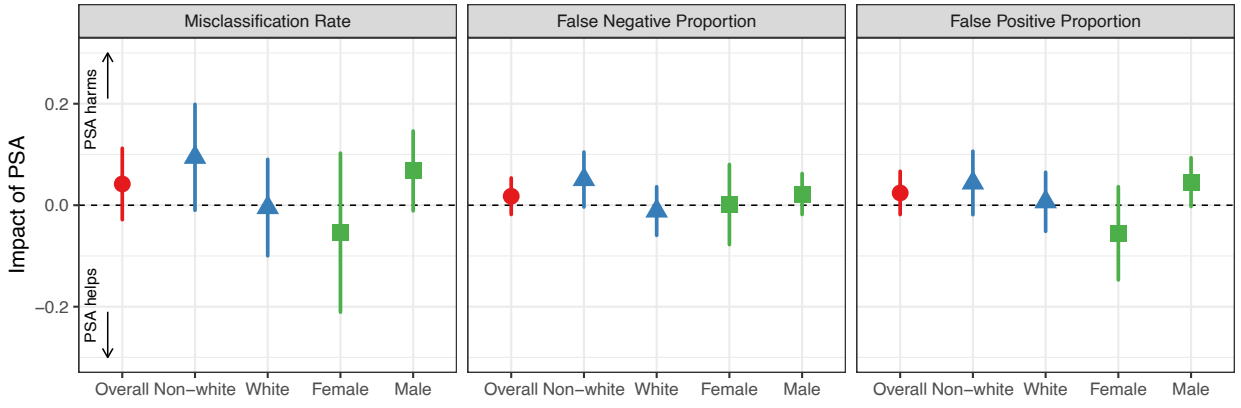

Fig. 1. Estimated impact of PSA recommendations on human decisions. The figure shows how PSA recommendations change a human judge's cash bail decisions in terms of misclassification rate, FNP, and FPP. The outcome variable is NCA. For each quantity of interest, we report a point estimate and its corresponding 95% CI for the overall sample (red circle), non-white and white subgroups (blue triangle), and female and male subgroups (green square). The results show that the PSA recommendations do not significantly improve the judge's decisions.

SI Appendix, Figs. S2 and S3 examine how often the judge correctly overrides the PSA recommendations by conducting a subgroup analysis based on whether or not the PSA recommends cash bail. For example, for the cases with $A_i = 1$, a true negative implies that the judge issues a signature bond decision against the PSA recommendation of cash bail to an arrestee who would not commit misconduct if released on their own recognizance. By comparing the true negative proportions between the human-alone and human-with-PSA system, we can adjust for the baseline disagreement between the human and PSA decisions. The estimates are qualitatively similar to those presented in Fig. 1, implying that the judge does not necessarily override the PSA recommendations correctly.

3.3. PSA-Alone Decisions Are Less Accurate Than Human Decisions. Next, we compare the classification performance of PSA-alone decisions with that of human decisions, using the proposed methodology described in Section 2.4. Specifically, we estimate the upper and lower bounds of the differences in the misclassification rate, FNP, and FPP between PSA-alone and human decisions (with and without PSA recommendations).

Fig. 2 shows that the PSA-alone system results in a substantially higher overall FPP, compared to the judge's decisions. The finding holds for non-white and male arrestees. For non-white arrestees, the misclassification rates are also significantly higher for the PSA-alone system than the human-alone system. For the FNP, the differences between the PSA-alone and human-alone systems are generally not statistically significant. This finding implies that the PSA system is generally harsher than the human judge, resulting in a greater number of unnecessary cash bail decisions across subgroups and different outcome variables. Similar results are obtained when comparing the PSA-alone and human-with-PSA systems (see *SI Appendix*, Fig. S4).

3.4. Human Decisions Are Preferred Over a PSA-Alone System When the Cost of False Positives Is High. Next, we analyze how one's loss function determines their preference over different decision-making systems. Specifically, we invert the hypothesis test using the bounds on the difference in classification risk derived in Theorem 3. This analysis allows us to estimate the range of the loss of FPs (ℓ_{01}), relative to the loss of FNs, which would lead us to prefer human decisions over the PSA-alone system.

We invert the hypothesis test shown in Eq. **2** over the range of values, $\ell_{01} \in [0.01, 100]$ using the 0.05 significance level. For each candidate value of ℓ_{01}, we conduct two one-sided hypothesis tests; one right-tailed and the other left-tailed, using the z-score of the lower and upper bounds of the difference in misclassification rates, respectively. If the left-tailed test null hypothesis, $H_0 : U_0 \geq 0$, is rejected (and thus the right-tailed test is not), the classification risk of the human-alone system is likely to be greater than that of PSA-alone system, suggesting a preference for the PSA-alone system over human decisions. Conversely, if the right-tailed test of null hypothesis $H_0 : L_0 \leq 0$ is rejected, it indicates a preference for the human-alone system. If neither test is rejected, the preference is ambiguous.

Fig. 3 shows that the human-alone system is preferred over the PSA-alone system when the loss of FP is about the same as or greater than that of FN. Overall, the human-alone system is preferred over the PSA-alone system when $\ell_{01} \geq 1.78$. Similar results are observed across various non-white and male subgroups. Exceptions are white and female arrestees, where we observe ambiguous results. Qualitatively similar results are also obtained when comparing the PSA-alone and human-with-PSA systems (see *SI Appendix*, Fig. S5), though we find that the results are ambiguous for white and male arrestees.

3.5. Optimally Combining PSA Recommendations with Human Decisions. Next, we investigate how to optimally integrate PSA recommendations into human decisions by applying the methods developed in Section 2.5. Specifically, we solve the empirical risk minimization problems outlined in Eqs. **3** and **4**, respectively. Here, we consider a policy class that maps FTA, NCA, and NVCA risk scores to a binary decision, subject to a monotonicity constraint (either increasing or decreasing). For example, under an increasing monotonicity constraint, if any of the three risk scores increases, the resulting binary rule should not decrease (i.e., a decision of $D = 1$ cannot then be altered to $D = 0$). We consider the NCA as the outcome.

The left plot of Fig. 4 shows when one should provide PSA recommendations to a human judge (indicated by dark blue) under this monotonicity constraint. We find that providing a PSA recommendation is advisable only when the FTA and NCA risk scores are relatively high and the NVCA flag is 1. The right plot shows when a human judge should follow the PSA recommendation (indicated by dark orange), again under the

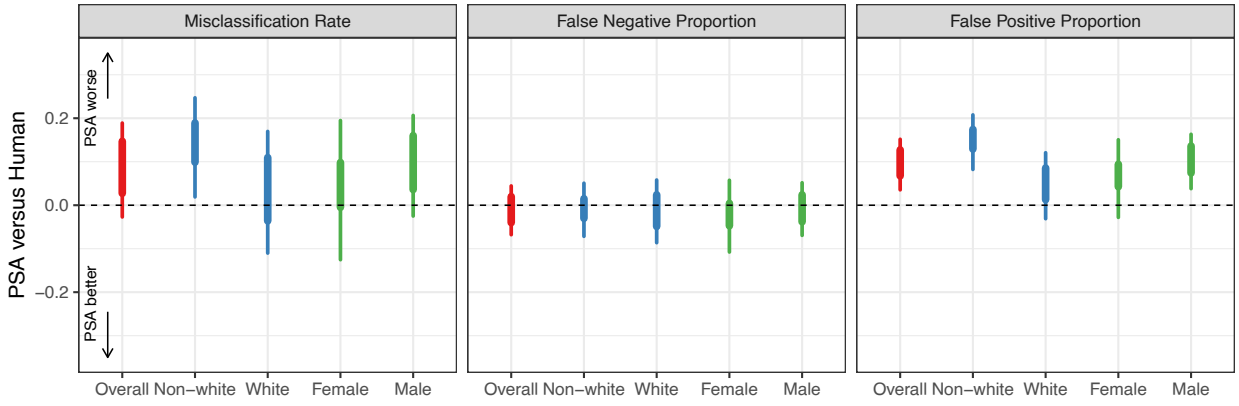

Fig. 2. Estimated bounds on difference in classification ability between PSA-alone and Human-alone decisions. The figure shows the misclassification rate, FNP, and FPP. The outcome variable is NCA. For each quantity of interest, we report estimated bounds (thick lines) and their corresponding 95% CI (thin lines) for the overall sample (red), non-white and white subgroups (blue), and female and male subgroups (green). The results indicate that PSA-alone decisions are less accurate than human judge's decisions in terms of the FPP.

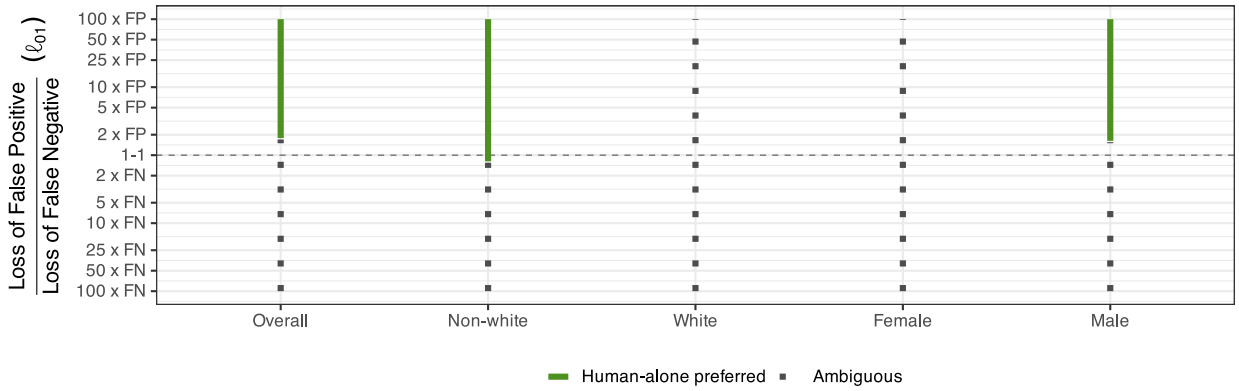

Fig. 3. Estimated preference for human-alone decisions over PSA-alone decision-making system. The figure illustrates the range of the ratio of the loss between false positives and false negatives, ℓ_{01}, for which one decision-making system is preferable over the other. A greater value of the ratio ℓ_{01} implies a greater loss of false positive relative to that of false negative. The figure displays the overall and subgroup-specific results. For each quantity of interest, we show the range of ℓ_{01} that corresponds to the preferred decision-making system; human-alone (green lines), and ambiguous (dotted lines). The results suggest that the human-alone system is preferred over the PSA-alone system when the loss of false positive is about the same as or greater than that of false negative. The PSA-alone system is never preferred within the specified range of ℓ_{01}.

monotonicity constraint. Our finding suggests that unless the FTA and NCA risk scores are relatively high and the NVCA flag is 1, a human judge should not follow the PSA recommendations and should instead use their judgment.

In sum, both the provision and decision rules suggest that one should not provide PSA input for the vast majority of cases. Rather, PSA should only be provided (or followed) in extreme cases where arrestees have many risk factors present. We emphasize that the magnitude of the improvement due to these optimal policies is small (see *SI Appendix*, Table S1). Our analysis shows that under the monotonicity constraint, the optimal provision of PSA recommendations results in a decrease of 0.01 in the misclassification rate when compared to not providing PSA recommendations for any case. Similarly, the optimal policy regarding when to follow PSA recommendations, under the monotonicity constraint, results in a decrease of 0.004 in the worst-case difference in the misclassification rate relative to not following PSA recommendations at all.

3.6. AI-Alone Decisions Are Less Accurate Than Human Decisions. It is natural to ask whether or not an alternative algorithmically generated risk score would perform better than the PSA. As an illustration, we compare the classification ability of AI-alone decisions with that of human decisions, using an open-source large language model, Llama3 (31) to generate AI decisions. We prompt the model to provide binary recommendations—

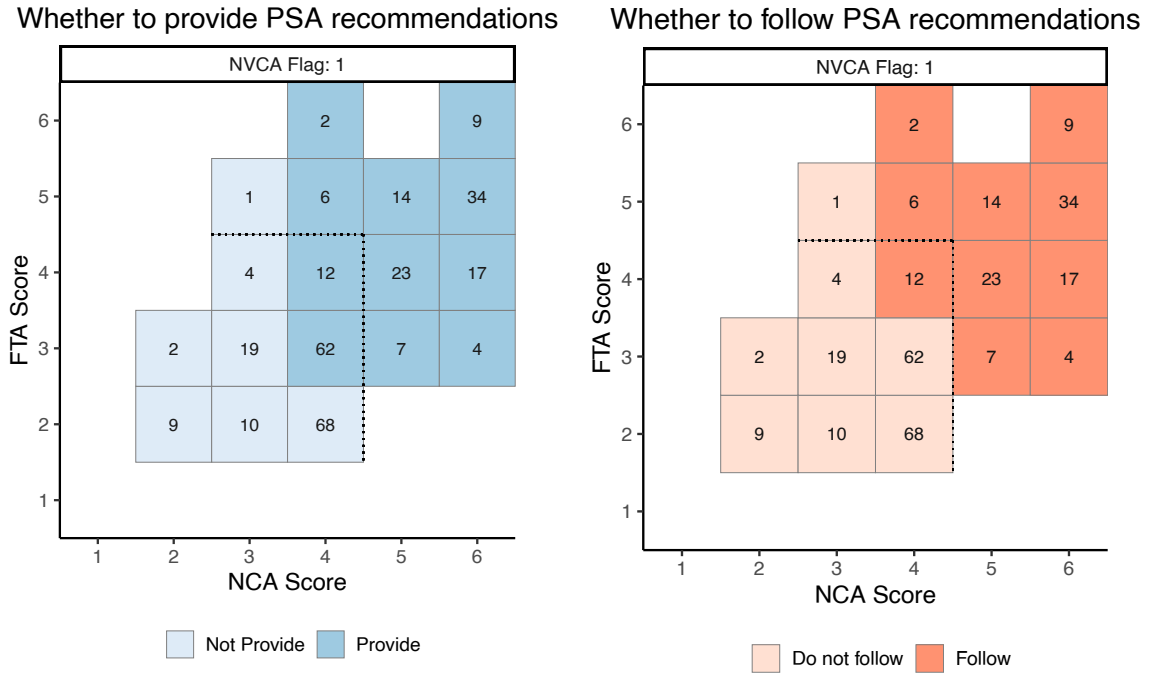

Fig. 4. Optimally Combining Human Decisions with PSA recommendations when NCA is the outcome. The left plot shows an estimated optimal policy for determining when to provide PSA recommendations to a human judge. The right plot shows an estimated optimal policy regarding when a human decision-maker should follow PSA recommendations. Each shaded area represents the optimal policy for specific combinations of risk scores: light shading indicates a decision rule of "not provide" (*Left*) or "do not follow" (*Right*), while dark shading indicates a decision rule of "provide" (*Left*) or "follow" (*Right*). Unshaded areas represent combinations of risk scores that are not possible. The number of observations for each combination is also shown.

whether to impose cash bail or release—based on the same set of PSA inputs for each arrestee. Here, we use deterministic decoding so that Llama3 always returns the same output for a given prompt (32). *SI Appendix*, Section S12 presents the prompt we use.

Using the methodology of Section 2.4, we estimate the upper and lower bounds of the differences in the misclassification rate, FNP, and FPP between Llama3 and human decisions. *SI Appendix*, Fig. S6 shows that the Llama3 decisions result in substantially higher FPP, compared to the judge's decisions. This finding holds across the overall sample and within every subgroup. The results suggest that the recommendations by Llama3 are generally harsher than the human judge, yielding a greater number of unnecessary cash bail decisions. For FNP, the differences between Llama3 and human decisions are generally not statistically significant.

4. Discussion

We have introduced a methodological framework for evaluating empirically the performance of three different decision-making systems: human-alone, AI-alone, and human-with-AI systems. We formalized the classification ability of each decision-making system using standard confusion matrices based on potential outcomes. We then showed that under single-blinded and unconfounded treatment assignment, we can directly identify the differences in classification ability between human decision-makers with and without AI recommendations. Furthermore, we derived partial identification bounds to compare the differences in classification ability between AI-alone and human decision-making systems and separately evaluate the performance of each system.

To illustrate the power of the proposed methodological framework, we applied our framework to the data from our own RCT whose goal is to evaluate the impact of the PSA risk assessment scores on a judge's decision to impose a cash bail or release arrestees on their own recognizance. We compared the human-alone and human-with-PSA decisions and found little to no impact of providing PSA recommendations. Our comparison of the human decision-maker with the PSA-alone system suggests, based on the baseline potential outcome and around 40% of the RCT's enrolled cases, that PSA-alone decisions may underperform as compared to human decisions, resulting in a greater proportion of unnecessarily harsh decisions. All together, these empirical findings suggest that integrating algorithmic recommendations into judicial decision-making warrants careful consideration and rigorous empirical evaluation.

There are several exciting future methodological research directions. The proposed methodological framework can be extended to common settings where decisions and outcomes are nonbinary (14). Another possible extension is the consideration of joint potential outcomes as done in ref. 1 and the dynamic settings where multiple decisions and outcomes are observed over time. Finally, the proposed methodology and its extensions can be applied to a variety of real-world settings where AI decision-making systems have been integrated or considered for future use.

Data, Materials, and Software Availability. Replication code and data have been deposited in Harvard Dataverse (https://doi.org/10.7910/DVN/KMM8WN) (16).

ACKNOWLEDGMENTS. D.J.G. and K.I. were partially supported by a NSF Grant (SES–2051196). Z.J. is partially supported by grants from the National Natural Science Foundation of China (Nos. 12371285 and 12292984).

Author affiliations: [a]Department of Statistics & Data Science and Heinz College of Information & Systems Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213; [b]Harvard Law School, Cambridge, MA 02138; [c]Department of Political Science, Yale University, New Haven, CT 06511; [d]Department of Statistics & Data Science, Yale University, New Haven, CT 06511; [e]Department of Government, Harvard University, Cambridge, MA 02138; [f]Department of Statistics, Harvard University, Cambridge, MA 02138; and [g]School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong 510275, China

Author contributions: E.B.-M., D.J.G., M.H., K.I., Z.J., and S.S. designed research; E.B.-M., D.J.G., M.H., K.I., Z.J., and S.S. performed research; E.B.-M., M.H., K.I., Z.J., and S.S. contributed new reagents/analytic tools; S.S. analyzed data; and E.B.-M., D.J.G., M.H., K.I., Z.J., and S.S. wrote the paper.

1. K. Imai, Z. Jiang, D. J. Greiner, R. Halen, S. Shin, Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *J. Royal Stat. Soc. Ser. A: Stat. Soc.* **186**, 167–189 (2023).
2. S. Barocas, M. Hardt, A. Narayanan, Fairness in machine learning. *Nips Tutorial* **1**, 2017 (2017).
3. S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv [Preprint] (2018). https://arxiv.org/abs/1808.00023 (Accessed 1 July 2025).
4. A. Chouldechova, A. Roth, A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* **63**, 82–89 (2020).
5. S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic fairness: Choices, assumptions, and definitions. *Annu. Rev. Stat. Its Appl.* **8**, 141–163 (2021).
6. K. Imai, Z. Jiang, Principal fairness for human and algorithmic decision-making. *Stat. Sci.* **38**, 317–328 (2023).
7. M. Hoffman, L. B. Kahn, D. Li, Discretion in hiring. *Q. J. Econ.* **133**, 765–800 (2018).
8. Y. Lai, A. Kankanhalli, D. Ong, "Human-AI collaboration in healthcare" in *Proceedings of the 54th Hawaii International Conference on System Sciences* (2021).
9. L. Cheng, A. Chouldechova, "Heterogeneity in algorithm-assisted decision-making: A case study in child abuse hotline screening" in *Proceedings of the ACM on Human-Computer Interaction* (2022), vol. 6, pp. 1–33.
10. J. Neyman, On the application of probability theory to agricultural experiments. Essay on principles. *Ann. Agric. Sci.* **5**, 465–472 (1923).
11. D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688 (1974).
12. H. Lakkaraju, J. Kleinberg, J. Leskovec, J. Ludwig, S. Mullainathan, "The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables" in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), pp. 275–284.
13. E. Chyn, B. Frandsen, E. C. Leslie, "Examiner and judge designs in economics: A practitioner's guide" (Working Paper No. 32348, National Bureau of Economic Research, 2024).
14. B. Koch, K. Imai, Statistical decision theory with counterfactual loss. arXiv [Preprint] (2025). https://arxiv.org/abs/2505.08908 (Accessed 1 July 2025).
15. D. J. Greiner, R. Halen, M. Stubenberg, J. Chistopher, L. Griffen, Randomized control trial evaluation of the implementation of the PSA-DMF system in Dane county (Tech. Rep., Access to Justice Lab, Harvard Law School, 2020).
16. E. Ben-Michael *et al.*, "Replication Data for: Does AI help humans make better decisions?: A statistical evaluation framework for experimental and observational studies." Harvard Dataverse. https://doi.org/10.7910/DVN/KMM8WN. Deposited 30 August 2025.
17. R. A. Berk, S. B. Sorenson, G. Barnes, Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *J. Empirical Legal Stud.* **13**, 94–115 (2016).
18. S. Goel, J. M. Rao, R. Shroff, Personalized risk assessments in the criminal justice system. *Am. Econ. Rev.* **106**, 119–123 (2016).
19. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018).
20. A. Rambachan, A. Coston, E. Kennedy, Counterfactual risk assessments under unmeasured confounding. arXiv [Preprint] (2022). https://arxiv.org/abs/2212.09844 (Accessed 1 July 2025).
21. D. Arnold, W. Dobbie, P. Hull, "Measuring racial discrimination in algorithms" in *AEA Papers and Proceedings* (American Economic Association, Nashville, TN, 2021), vol. 111, pp. 49–54.
22. D. Arnold, W. Dobbie, P. Hull, Measuring racial discrimination in bail decisions. *Am. Econ. Rev.* **112**, 2992–3038 (2022).
23. R. Binns *et al.*, "'It's reducing a human being to a percentage' perceptions of justice in algorithmic decisions" in *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems* (2018), pp. 1–14.
24. C. J. Cai, S. Winter, D. Steiner, L. Wilcox, M. Terry, "hello AI": Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making" in *Proceedings of the ACM on Human-Computer Interaction* (2019), vol. 3, pp. 1–24.
25. E. Ben-Michael, K. Imai, Z. Jiang, Policy learning with asymmetric counterfactual utilities. *J. Am. Stat. Assoc.* **119**, 3045–3058 (2024).
26. J. M. Robins, A. Rotnitzky, L. P. Zhao, Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* **89**, 846–866 (1994).
27. J. Y. Audibert, A. B. Tsybakov, Fast learning rates for plug-in classifiers. *Ann. Stat.* **35**, 608–633 (2007).
28. G. W. Imbens, C. F. Manski, Confidence intervals for partially identified parameters. *Econometrica* **72**, 1845–1857 (2004).
29. S. Athey, S. Wager, Policy learning with observational data. *Econometrica* **89**, 133–161 (2021).
30. M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics* (Cambridge University Press, 2019).
31. H. Touvron *et al.*, Llama: Open and efficient foundation language models. arXiv [Preprint] (2023). https://arxiv.org/abs/2302.13971 (Accessed 1 July 2025).
32. K. Imai, K. Nakamura, Causal representation learning with generative artificial intelligence: Application to texts as treatments. arXiv [Preprint] (2024). https://arxiv.org/abs/2410.00903 (Accessed 1 July 2025).