Supporting Information for

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

Eli Ben-Michael, D. James Greiner, Melody Huang, Kosuke Imai, Zhichao Jiang, and Sooahn Shin

Kosuke Imai.
E-mail: imai@harvard.edu

This PDF file includes:

Supporting Information Text

S1. Contributions to the Literature

Our work contributes to a growing literature that addresses the selective labels problem when evaluating human decisions and AI recommendations. In particular, we consider an evaluation design in which the provision of AI recommendations is either randomized or assumed to be unconfounded while single-blinding the treatment assignment so that AI recommendations affect the outcome only through human decisions. We show that under this design, it is possible to evaluate the classification performance of human-alone, AI-alone, and human-with-AI decision-making systems. In contrast, related studies have been restricted to observational settings. For example, previous works exploit discontinuities at algorithmic thresholds and staggered roll-outs of algorithms (e.g. 1–5) or use survey evaluations (e.g., 6, 7).

Several studies have advocated designs that use quasi-random assignment of cases to different decision-makers. They use the differing decision rates as an instrumental variable to estimate various performance measures of AI recommendations and/or human decision-makers (e.g. 8–11). Unlike these studies, which rely on two-stage least squares, our approach does not assume monotonicity (though their causal quantities of interest differ from ours). As a result, under the proposed experimental setting, we can guarantee the required identification assumptions *by design*. Furthermore, while related approaches have been used to evaluate algorithmic decisions, we also evaluate the relative performance of human decision-makers with and without AI recommendations, as well as the AI-alone decision-making system.

Closest to our approach is (11), who compare AI-assisted human decisions to those of the algorithm by studying cases where humans override the algorithmic recommendation (see also 12). Our framework is also similar to the one proposed by (13), but we focus on a single potential outcome rather than joint potential outcomes, allowing us to avoid making additional assumptions.

When point identification is not possible, we use partial identification to bound the quantities of interest (e.g., 14). This methodological development is related to partial identification approaches proposed by (15) and (16). In particular, (16) consider general approaches to partial identification of the predictive performance of classification algorithms. In contrast, we focus on comparing the predictive performance of the aforementioned three different decision-making systems that involve humans and/or AI, leading to different identification results and estimation strategies.

Prior work has also considered classification ability measures that are related to ours. For example, in the pre-trial risk assessment setting, (11) consider the misconduct rate among released defendants; this corresponds to the *false negative rate*. (9) consider the proportion of individuals who are detained erroneously, i.e., the *false discovery rate*. Other work develops a more general framework. For instance, (16) consider a generalized notion of performance that includes functions of the confusion matrix as well as other measures like calibration and mean square error, which we do not consider here.

In contrast to these approaches, (13) introduce a principal stratification framework that considers the joint set of potential outcomes and three principal strata of individuals: (1) *preventable cases* $(Y(1), Y(0)) = (0, 1))$ — individuals who would engage in misconduct only if released, (2) risky cases $(Y(1), Y(0)) = (1, 1))$ — individuals who would engage in misconduct regardless of the judge's decision, and (3) safe cases $(Y(1), Y(0) = (0, 0))$ — individuals who would not engage in misconduct regardless of the detention decision. The authors focus on the effect of AI recommendation provision on the human decision, conditioned on these principal strata.

(17) also introduce a related fairness notion, called principal fairness. These quantities—although relying on both potential outcomes—can be related to the loss function perspective we take. In particular, under the strong assumption that a positive decision entirely prevents a positive outcome (i.e. $Y(1) = 0$ for all individuals), there are only preventable and safe cases. In that case, the effect on the judge's decision given that a case is preventable is the effect on the false positive rate, while the effect on the judge's decision given that a case is safe is the effect on the false negative rate. Whether such a connection exists without the strong assumption that $Y(1) = 0$ is an open question.

Our estimation strategy draws heavily on principles of doubly-robust estimation in randomized control trials and observational studies that leverage estimates of both the propensity score and outcome model to efficiently estimate treatment effects. See (18–20) for recent reviews and perspectives on the long literature on this subject. In addition, our estimation strategy for upper and lower bounds draws on extensions of the double-robust methodology to partially identified parameters. Examples of such work can be found in (21–24).

Finally, our empirical results clarify whether providing judges with the PSA at the predisposition stage improves criminal justice outcomes. This is a question that has thus far lacked consistent evidence, both for the PSA specifically and for risk assessment tools more broadly. Three unpublished before-and-after studies of reform packages that included PSA adoption have reported conflicting findings. Two found statistically significant increases in the use of own recognizance release by 20 and 11 percentage points, respectively, without corresponding changes in failure to appear (FTA), new criminal activity (NCA), or, in the only study to measure it, the number of days spent in predisposition detention (25, 26). A third study reported statistically significant reductions in FTA (from 30% to 24%), NCA (20% to 15%), and new violent criminal activity (NVCA, 6% to 4%) following implementation of a similar PSA-based reform (27). A fourth study observed a temporary increase in own-recognizance decisions, a longer-term rise in FTA, and no meaningful changes in NCA or NVCA (28).

Furthermore, the only randomized controlled trial (RCT) of a bail-stage risk assessment instrument that we are aware of — and indeed, the first RCT ever conducted in the legal field — found that providing a simple risk tool, along with a structured phone reminder program, increased the share of defendants released on their own recognizance from 14% to 60%, with no statistically significant change in FTA (29).

S2. Generic Loss Functions

As a generic loss function, we can define separate weights for true positives ℓ_{11}, true negatives ℓ_{00} and false positive ℓ_{01} so that the expected loss is given by $R(\ell_{00}, \ell_{01}, \ell_{11}) = \ell_{10}q_{10} + \ell_{01}q_{01} + \ell_{11}q_{11} + \ell_{00}q_{00}$ with a proper normalization constraint such as $\ell_{10} = 1$. All of the quantities we have considered in the main text are special cases of this generic loss function. For example, the difference in risk between the human-with-AI and human-alone systems is given by,

$$\begin{aligned} & R_{\text{HUMAN+AI}}(\ell_{00}, \ell_{01}, \ell_{11}) - R_{\text{HUMAN}}(\ell_{00}, \ell_{01}, \ell_{11}) \\ &= p_{10}(D(1)) - p_{10}(D(0)) + \ell_{01}(p_{01}(D(1)) - p_{01}(D(0))) + \ell_{11}\left(p_{11}(D(1)) - p_{11}(D(0))\right) \\ &\quad + \ell_{00}\left(p_{00}(D(1)) - p_{00}(D(0))\right). \end{aligned}$$

Now, recall that although the false positive proportion under each system $p_{01}(D(z))$ is not identified, the difference between $p_{01}(D(1))$ and $p_{01}(D(0))$ is identifiable as $p_{01}(D(1)) - p_{01}(D(0)) = p_{00}(D(0)) - p_{00}(D(1))$. We can similarly point identify the difference in true positive proportions as $p_{11}(D(1)) - p_{11}(D(0)) = p_{10}(D(0)) - p_{10}(D(1))$. So, following the argument for Theorem 1, we can point identify the difference in risk with the generic loss function as:

$$\begin{aligned} & R_{\text{HUMAN+AI}}(\ell_{00}, \ell_{01}, \ell_{11}) - R_{\text{HUMAN}}(\ell_{00}, \ell_{01}, \ell_{11}) \\ &= (1 - \ell_{11})\left(p_{10}(D(1)) - p_{10}(D(0))\right) + (\ell_{00} - \ell_{01})\left(p_{00}(D(1)) - p_{00}(D(0))\right). \end{aligned}$$

We can similarly evaluate the human-with-AI vs AI-alone and human-alone vs AI-alone systems by following the partial identification arguments in Section D, and evaluate the decision-making systems independently directly using the results shown in Section S10.

S3. Classification Risk of a Generic Decision-making System

In Section D, we focus on the evaluation of an AI-alone decision-making system based on the specific AI recommendation A used in the experiment. For completeness, here we show how to evaluate the classification ability of any alternative decision-making system D^* in itself that satisfies the following conditional independence $D^* \perp\!\!\!\perp Y(0) \mid A, X$. Define such a decision-making system as $f(a, x) := \Pr(D^* = 1 \mid A = a, X = x)$. The classification risk of this decision-making system D^* can be written as,

$$R(\ell_{01}; D^*) \quad = \quad \mathbb{E}\left[(1 - f(A, X)) \Pr(Y(0) = 1 \mid A, X) + \ell_{01} f(A, X) \Pr(Y(0) = 0 \mid A, X)\right].$$

The sharp bounds are given by the following theorem.

THEOREM S1 (SHARP BOUNDS ON THE CLASSIFICATION RISK OF A GENERIC STOCHASTIC DECISION-MAKING SYSTEM) *Consider the decision making system $f(a, x) := \Pr(D^* = 1 \mid A = a, X = x)$ that satisfies the conditional independence relation, $D^* \perp\!\!\!\perp Y(0) \mid X, A$. The sharp bounds on its classification risk $R(\ell_{01}; D^*)$ are given by:*

$$\begin{aligned} R(\ell_{01}; D^*) \quad \in \quad & \Bigg[\mathbb{E}\Big[\ell_{01} \cdot f(A, X) + \{1 - (1 + \ell_{01}) f(A, X)\}\left[g_f(A, X) \max_{z'} \Pr(Y = 1, D = 0 \mid A, X, Z = z')\right. \\ & \qquad \left. + (1 - g_f(A, X))\{1 - \max_{z'} \Pr(Y = 0, D = 0 \mid A, X, Z = z')\}\right]\Big], \\ & \quad \mathbb{E}\Big[\ell_{01} \cdot f(A, X) + \{1 - (1 + \ell_{01}) f(A, X)\}\left[g_f(A, X)\{1 - \max_{z'} \Pr(Y = 0, D = 0 \mid A, X, Z = z')\}\right. \\ & \qquad \left. + (1 - g_f(A, X)) \max_{z'} \Pr(Y = 1, D = 0 \mid A, X, Z = z')\right]\Big]\Bigg] \end{aligned}$$

where $g_f(a, x) = \mathbb{1}\{1 - (1 + \ell_{01}) f(a, x) \geq 0\}$.

S4. Technical Assumptions

ASSUMPTION S1 For each $z = 0, 1$, we have:

$$\begin{aligned} & \left(\|m^D(z, \cdot) - \hat{m}^D(z, \cdot)\|_2 + \|m^Y(z, \cdot) - \hat{m}^Y(z, \cdot)\|_2\right) \times \|e - \hat{e}\|_2 = o_p(n^{-\frac{1}{2}}), \\ & \|m^Y(z, \cdot) - \hat{m}^Y(z, \cdot)\|_\infty \ = o_p(1), \quad \|m^D(z, \cdot) - \hat{m}^D(z, \cdot)\|_\infty = o_p(1), \quad \|e - \hat{e}\|_\infty = o_p(1), \end{aligned}$$

where for a given function f, $\|f\|_2^2 = \mathbb{E}[f(X)^2]$ and $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$.

Assumption S1 relates to the standard product-rate assumption for doubly-robust AIPW estimators, but it involves both the decision and outcome models because the compound outcome W_i involves the product of the decision and the outcome.* In a randomized experiment, the propensity scores are known, and so we only require that we consistently estimate the outcome models, with no particular rate requirement.

*To establish Theorem 2, it is sufficient to have only a rate requirement for a combination of the two models. For clarity, however, we give a somewhat stronger sufficient condition in Assumption S1.

Assumption S2 There exist constants $C > 0$ and $\alpha > 0$ such that:

$$\Pr\left(\left|(1 - m^D(1 - z, X, 0))m^Y(1 - z, x, 0) - (1 - m^D(z, x, 0))m^Y(z, x, 0)\right| \le t\right) \;\le\; Ct^{\alpha},$$
$$\Pr\left(\left|(1 - m^D(1 - z, X, 0))(1 - m^Y(1 - z, x, 0)) - (1 - m^D(z, x, 0))(1 - m^Y(z, x, 0))\right| \le t\right) \;\le\; Ct^{\alpha}.$$

Larger values of the margin parameter α imply that the difference in the bounds is often large, and so it is easy to classify which is tighter. Conversely, smaller values of α mean that the classification problem is harder because the difference between the bounds is often small. In the continuous case, if the covariates have a bounded density, then $\alpha >= 1$ (30). Margin conditions such as this have been used when estimating partially identified parameters (21, 31–33) and for policy learning (23, 24, 34, 35). This margin condition, along with the following additional rate conditions, establish the asymptotical normality of the estimated bounds.

Assumption S3 For $z = 0, 1$ and $a = 0, 1$,

1. $\left(\|m^Y(z, \cdot, 0) - \hat{m}^Y(z, \cdot, 0)\|_2 + \|m^D(z, \cdot, a) - \hat{m}^D(z, \cdot, a)\|_2\right) \times \|\hat{e}(z, \cdot) - e(z, \cdot)\|_2 = o_p\left(n^{-1/2}\right)$, $\|m^Y(z, \cdot, 0) - \hat{m}^Y(z, \cdot, 0)\|_\infty = o_p(1)$, *and* $\|m^D(z, \cdot, a) - \hat{m}^D(z, \cdot, a)\|_\infty = o_p(1)$
2. $(\|\hat{m}^D(z, \cdot, 0) - m^D(z, \cdot, 0)\|_\infty + \|\hat{m}^Y(z, \cdot, 0) - m^Y(z, \cdot, 0)\|_\infty)^{1+\alpha} = o_p\left(n^{-\frac{1}{2}}\right)$

The first rate condition in Assumption S3 is analogous to the rate condition in Assumption S1, but for the outcome and decision models conditional on the AI recommendation A. As before, in a randomized experiment, we only require consistency of the outcome and decision model estimates.

In contrast, the second condition in Assumption S3 requires that we can estimate the outcome and decision models at a sufficiently fast rate to estimate the nuisance classifiers well. The margin parameter α determines how fast the rate needs to be. If the classification task is more difficult and α is small, then we will need to estimate the nuisance components at closer to the parametric $n^{-1/2}$ rate; if the task is easier and α is large, then the rate can be slower. However, the required rate is always strictly slower than the parametric rate because $\alpha > 0$. The knowledge of propensity score in a randomized experiment does not remove this requirement, though we can choose what covariates to include. Including more covariates can lead to more informative bounds, although estimation may become more challenging.

S5. Exact Expressions of the Sharp Bounds of Theorem 3

$$\begin{aligned}
L_z(x) \;&:=\; (1 + \ell_{01})\left\{\max_{z'} \Pr(Y = 1, D = 0, A = 0 \mid Z = z', X = x) - \Pr(Y = 1, D = 0 \mid Z = z, X = x)\right\} \\
&\quad + \ell_{01}\left\{\Pr(D = 0, A = 1 \mid Z = z, X = x) - \Pr(D = 1, A = 0 \mid Z = z, X = x)\right\}, \\
U_z(x) \;&:=\; (1 + \ell_{01})\Bigg\{\Pr(A = 0 \mid X = x) - \Pr(Y = 1, D = 0 \mid Z = z, X = x) \\
&\qquad - \max_{z'} \Pr(Y = 0, D = 0, A = 0 \mid Z = z', X = x)\Bigg\} \\
&\quad + \ell_{01}\left\{\Pr(D = 0, A = 1 \mid Z = z, X = x) - \Pr(D = 1, A = 0 \mid Z = z, X = x)\right\}.
\end{aligned}$$

S6. Efficient Estimators of the Sharp Bounds

The efficient estimators of the sharp bounds are based on the following two sets of efficient estimators. The first is for the models of $Y(1 - D)(1 - A)$ and $(1 - Y)(1 - D)(1 - A)$,

$$\begin{aligned}
\widehat{\varphi}_{z1}(Z, X, D, A, Y) &= (1 - A)(1 - \hat{m}^D(z, X, 0))\hat{m}^Y(z, X, 0) + \frac{\mathbb{1}\{Z = z\}(1 - A)(1 - D)}{\hat{e}(z, X)}(Y - \hat{m}^Y(z, X, 0)) \\
&\quad - \hat{m}^Y(z, X, 0)\frac{\mathbb{1}\{Z = z\}(1 - A)}{\hat{e}(z, X)}(D - \hat{m}^D(z, X, 0)), \\
\widehat{\varphi}_{z0}(Z, X, D, A, Y) &= (1 - A)(1 - \hat{m}^D(z, X, 0))(1 - \hat{m}^Y(z, X, 0)) - \frac{\mathbb{1}\{Z = z\}(1 - A)(1 - D)}{\hat{e}(z, X)}(Y - \hat{m}^Y(z, X, 0)) \\
&\quad - (1 - \hat{m}^Y(z, X, 0))\frac{\mathbb{1}\{Z = z\}(1 - A)}{\hat{e}(z, X)}(D - \hat{m}^D(z, X, 0)).
\end{aligned}$$

The second set is for the models of $A(1 - D)$, and $(1 - A)D$,

$$\begin{aligned}
\widehat{\varphi}^D_{z1}(Z, X, D, A) &= A(1 - \hat{m}^D(z, X, 1)) - \frac{\mathbb{1}\{Z = z\}}{\hat{e}(Z, X)}A(D - \hat{m}^D(z, X, 1))), \\
\widehat{\varphi}^D_{z0}(Z, X, D, A) &= (1 - A)\hat{m}^D(z, X, 0) + \frac{\mathbb{1}\{Z = z\}}{\hat{e}(Z, X)}(1 - A)(D - \hat{m}^D(z, X, 0)).
\end{aligned}$$

As before, we remove the circumflexes to refer to the true uncentered influence functions. Finally, we estimate the upper and lower bound as

$$\begin{aligned}
\widehat{L}_z = \frac{1}{n}\sum_{i=1}^{n} &(1+\ell_{01})\left(\widehat{\varphi}_{z1}(Z_i,X_i,D_i,A_i,Y_i) - \widehat{\varphi}_z(Z_i,X_i,D_i,Y_i;0)\right)\\
&+ \ell_{01}\left(\widehat{\varphi}^D_{z1}(Z_i,X_i,D_i,A_i,Y_i) - \widehat{\varphi}^D_{z0}(Z_i,X_i,D_i,A_i,Y_i)\right)\\
&+ (1+\ell_{01})\hat{g}_{L_z}(X_i)\left(\widehat{\varphi}_{1-z,1}(Z_i,X_i,D_i,A_i,Y_i) - \widehat{\varphi}_{z1}(Z_i,X_i,D_i,A_i,Y_i)\right),\\
\widehat{U}_z = \frac{1}{n}\sum_{i=1}^{n} &(1+\ell_{01})\left(\widehat{\varphi}_{z1}(Z_i,X_i,D_i,A_i,Y_i) - \widehat{\varphi}_z(Z_i,X_i,D_i,Y_i;0)\right)\\
&+ \ell_{01}\widehat{\varphi}^D_{z1}(Z_i,X_i,D_i,A_i,Y_i) + \widehat{\varphi}^D_{z0}(Z_i,X_i,D_i,A_i,Y_i)\\
&- (1+\ell_{01})\hat{g}_{U_z}(X_i)\left(\widehat{\varphi}_{1-z,0}(Z_i,X_i,D_i,A_i,Y_i) - \widehat{\varphi}_{z0}(Z_i,X_i,D_i,A_i,Y_i)\right).
\end{aligned}$$

S7. Exact Expressions of the Asymptotic Variances of Theorem 4

$$\begin{aligned}
V_{L_z} &= \mathbb{E}[\{(1+\ell_{01})(\varphi_z(Z,X,D,A,Y) - \varphi_z(Z,X,D,Y;0)) + \ell_{01}(\varphi^D_{z1}(Z,X,D,A,Y) - \varphi^D_{z0}(Z,X,D,A,Y))\\
&\qquad + (1+\ell_{01})g_{L_z}(X)\left(\varphi_{1-z,1}(Z,X,D,A,Y) - \varphi_z(Z,X,D,A,Y)\right) - L_z\}^2],\\
V_{U_z} &= \mathbb{E}[\{(1+\ell_{01})\varphi_{z1}(Z,X,D,A,Y) - \varphi_z(Z,X,D,Y;0) + \ell_{01}\varphi^D_{z1}(Z,X,D,A,Y) + \varphi^D_{z0}(Z,X,D,A,Y)\\
&\qquad\qquad + (1+\ell_{01})g_{U_z}(X)\left(\varphi_{1-z,0}(Z,X,D,A,Y) - \varphi_{z0}(Z,X,D,A,Y)\right)\}^2].
\end{aligned}$$

S8. Minimizing the Excess Worst-case Risk

We take an empirical risk minimization approach and find a policy $\hat{\pi}_{\text{DEC}}$ that minimizes our estimate of the worst-case excess risk by solving:

$$\begin{aligned}
\hat{\pi}_{\text{DEC}} \in \operatorname*{arg\,min}_{\pi\in\Pi}\frac{1}{n}\sum_{i=1}^{n}\pi(X_i)\,&[(1+\ell_{01})\left(\widehat{\varphi}_{01}(Z_i,X_i,D_i,A_i,Y_i) - \widehat{\varphi}_0(Z_i,X_i,D_i,Y_i;0)\right)\\
&+ \ell_{01}\widehat{\varphi}^D_{01}(Z_i,X_i,D_i,A_i,Y_i) + \widehat{\varphi}^D_{00}(Z_i,X_i,D_i,A_i,Y_i)\\
&-(1+\ell_{01})\hat{g}_{U_0}(X_i)\left(\widehat{\varphi}_{10}(Z_i,X_i,D_i,A_i,Y_i) - \widehat{\varphi}_{00}(Z_i,X_i,D_i,A_i,Y_i)\right)].
\end{aligned}\tag{S1}$$

Due to the lack of point identification, we bound the excess *worst-case* risk of the estimated policy $\hat{\pi}_{\text{DEC}}$ versus the population policy π^*_{DEC}.

THEOREM S2 *Under Assumptions 1, S1, S2, and S3, we have:*

$$\begin{aligned}
&\mathbb{E}[(\hat{\pi}_{\text{DEC}}(X) - \pi^*_{\text{DEC}})U_z(X)]\\
&\leq C\Bigg(\sum_{z'=0}^{1}\|m^Y(z',\cdot,0) - \hat{m}^Y(z',\cdot,0)\|_2 + \|m^D(z',\cdot,0) - \hat{m}^D(z',\cdot,0)\|_2\\
&\quad + \quad \|m^Y(z,\cdot) - \hat{m}^Y(z,\cdot)\|_2 + \|m^D(z,\cdot) - \hat{m}^D(z,\cdot)\|_2 + \|m^D(z,\cdot,1) - \hat{m}^D(z,\cdot,1)\|_2\Bigg)\\
&\qquad \times \|\hat{e} - e\|_2 + 2C(\|\hat{m}^D(\cdot,\cdot,0) - m^D(\cdot,\cdot,0)\|_\infty + \|\hat{m}^Y(\cdot,\cdot,0) - m^Y(\cdot,\cdot,0)\|_\infty)^{1+\alpha}\\
&\quad + \left(1+\frac{2}{\eta}\right)(4+6\ell_{01})\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}},
\end{aligned}$$

with probability at least $1 - 2\exp(-t^2/2)$.

Theorem S2 shows that the error in the nuisance components and the complexity of the policy class control the excess worst-case risk. In contrast to the case of the estimated AI-recommendation provision rule $\hat{\pi}_{\text{REC}}$, however, the knowledge of the propensity score is not sufficient for the estimated AI-alone decision rule $\hat{\pi}_{\text{REC}}$ to have low excess risk. As with Theorem 4, because optimizing for the worst-case excess risk involves estimating sharp bounds (and the nuisance classifier $g_{U_z}(x)$), the estimation error of the outcome and decision models enter into the bound alone.

S9. Lemmas

We present two lemmas used to derive sharp bounds on classification risks and their differences. These lemmas provide sharp bounds on the two key unidentifiable quantities $\theta_a := \Pr(Y(0) = 1, D = 1, A = a)$ and $\xi_{az} := \Pr(Y(0) = 1, D(z) = 1, A = a)$. For notational simplicity, we omit covariates. However, the two lemmas continue to hold if we condition on covariates $\mathbf{X}$.

Lemma 1 *Define $\theta_a := \Pr(Y(0) = 1, D = 1, A = a)$ for $a = 0, 1$. Then, under Assumption 1, its sharp bounds are given by $\theta_a \in [\underline{\theta}_a, \bar{\theta}_a]$ where*

$$\begin{aligned}
\underline{\theta}_a &= \max_z \Pr(Y = 1, D = 0, A = a \mid Z = z) - \Pr(Y = 1, D = 0, A = a), \\
\bar{\theta}_a &= \Pr(A = a) - \Pr(Y = 1, D = 0, A = a) - \max_z \Pr(Y = 0, D = 0, A = a \mid Z = z).
\end{aligned}$$

for $a = 0, 1$. These sharp bounds can be achieved simultaneously for $a = 0, 1$.

Lemma 2 *Define $\xi_{az} := \Pr\{Y(0) = 1, D(z) = 1, A = a\}$ for $a, z = 0, 1$. Then, under Assumption 1, its sharp bounds are given by $\xi_{az} \in [\underline{\xi}_{az}, \bar{\xi}_{az}]$ where*

$$\begin{aligned}
\underline{\xi}_{az} &= \max_{z'} \Pr(Y = 1, D = 0, A = a \mid Z = z') - \Pr(Y = 1, D = 0, A = a \mid Z = z), \\
\bar{\xi}_{az} &= \Pr(A = a) - \Pr(Y = 1, D = 0, A = a \mid Z = z) - \max_{z'} \Pr(Y = 0, D = 0, A = a \mid Z = z')
\end{aligned}$$

for $a, z = 0, 1$. These sharp bounds can be achieved simultaneously for $a = 0, 1$ given $z = 0, 1$.

S10. Separate Evaluation of Each Decision-making System

While we have focused on identifying and bounding the *differences* in classification risks of the three decision-making systems — human-alone, AI-alone, and human-with-AI, it is also possible to partially identify the classification risk of each decision-making system separately.

Theorem S3 (Sharp bounds on the classification risk of each decision-making system) *The sharp bounds on the risk of each decision-making system are given by:*

(a) Human-alone system:

$$R_{\textsc{Human}}(\ell_{01}) \in [p_{10}(D(0)) + \ell_{01} \cdot \underline{p}_{01}(D(0)),\ p_{10}(D(0)) + \ell_{01} \cdot \bar{p}_{01}(D(0))],$$

(b) Human-with-AI system

$$R_{\textsc{Human+AI}}(\ell_{01}) \in [p_{10}(D(1)) + \ell_{01} \cdot \underline{p}_{01}(D(1)),\ p_{10}(D(1)) + \ell_{01} \cdot \bar{p}_{01}(D(1))],$$

(c) AI-alone system:

$$\begin{aligned}
R_{\textsc{AI}}(\ell_{01}) \quad \in \quad & \Big[\max_{z'} \Pr(Y = 1, D = 0, A = 0 \mid Z = z') + \ell_{01} \cdot \max_{z'} \Pr(Y = 0, D = 0, A = 1 \mid Z = z')\ , \\
& \Pr(A = 0) - \max_{z'} \Pr(Y = 0, D = 0, A = 0 \mid Z = z') \\
& + \ell_{01} \cdot \Big\{\Pr(A = 1) - \max_{z'} \Pr(Y = 1, D = 0, A = 1 \mid Z = z')\Big\}\Big],
\end{aligned}$$

where $\underline{p}_{01}(D(z)) := \Pr(D = 1 \mid Z = z) - \bar{\xi}_{0z} - \bar{\xi}_{1z}$ and $\bar{p}_{01}(D(z)) := \Pr(D = 1 \mid Z = z) - \underline{\xi}_{0z} - \underline{\xi}_{1z}$ with the following upper and lower bounds of $\xi_{az} := \Pr\{Y(0) = 1, D(z) = 1, A = a\}$:

$$\begin{aligned}
\underline{\xi}_{az} &= \max_{z'} \Pr(Y = 1, D = 0, A = a \mid Z = z') - \Pr(Y = 1, D = 0, A = a \mid Z = z), \\
\bar{\xi}_{az} &= \Pr(A = a) - \Pr(Y = 1, D = 0, A = a \mid Z = z) - \max_{z'} \Pr(Y = 0, D = 0, A = a \mid Z = z')
\end{aligned}$$

for $a, z = 0, 1$.

Similarly, we can derive the partial identification bounds for non-linear classification measures, such as FNR, FPR, and FDR.

S11. Proofs

A. Proof of Theorem 1. The law of total probability implies:

$$\Pr(Y(0)=y) \;=\; p_{y1}(D(z))+p_{y0}(D(z)), \text{ for } z\in\{0,1\}.$$

Then, we have:

$$p_{y1}(D(1))+p_{y0}(D(1)) = p_{y1}(D(0))+p_{y0}(D(0)) \implies p_{y1}(D(1))-p_{y0}(D(0)) = p_{y0}(D(0))-p_{y1}(D(1)),$$

for $y\in\{0,1\}$. Using this result, we obtain:

$$\begin{aligned} R_{\text{HUMAN+AI}}(\ell_{01})-R_{\text{HUMAN}}(\ell_{01}) &= p_{10}(D(1))+\ell_{01}p_{01}(D(1))-\{p_{10}(D(0))+\ell_{01}p_{01}(D(0))\} \\ &= p_{10}(D(1))-p_{10}(D(0))+\ell_{01}\left\{p_{01}(D(1))-p_{01}(D(0))\right\} \\ &= p_{10}(D(1))-p_{10}(D(0))-\ell_{01}\left\{p_{00}(D(1))-p_{00}(D(0))\right\}. \end{aligned} \qquad \text{[S2]}$$

Under Assumption 1, we have

$$\begin{aligned} p_{y0}(D(z)) &= \mathbb{E}[\Pr\{Y(0)=y, D(z)=0 \mid X\}] \\ &= \mathbb{E}[\Pr\{Y(0)=y, D(z)=0 \mid Z=z, X\}] \\ &= \mathbb{E}[\Pr\{Y=y, D=0 \mid Z=z, X\}]. \end{aligned} \qquad \text{[S3]}$$

Plugging Eq. (S3) into Eq. (S2) yields the identification formula. □

B. Proof of Theorem 2. Define β_z and then rewrite it as

$$\begin{aligned} \beta_z &= \mathbb{E}[\Pr(Y=1, D=0 \mid Z=z, X)-\ell_{01}\times\Pr(Y=0, D=0 \mid Z=z, X)] \\ &= \mathbb{E}[(1-m^D(z,X))m^Y(z,X)-\ell_{01}\times(1-m^D(z,X))(1-m^Y(z,X))] \\ &= \mathbb{E}[(1-m^D(z,X))\{(1+\ell_{01})m^Y(z,X)-\ell_{01}\}] \\ &= \mathbb{E}[(1-D(z))\{(1+\ell_{01})Y(z)-\ell_{01}\}] \\ &= \mathbb{E}[\mathbb{E}[W \mid X, Z=z]]. \end{aligned}$$

Using this, we can write the classification risk difference as $R_{\text{HUMAN+AI}}(\ell_{01})-R_{\text{HUMAN}}(\ell_{01})=\beta_1-\beta_0$.

Next, we define:

$$m(z,x;\ell_{01}) := \mathbb{E}[W \mid X=x, Z=z] = \mathbb{E}[(1-D)\{(1+\ell_{01})Y-\ell_{01}\} \mid X=x, Z=z].$$

Then, the (uncentered) efficient influence function is given by,

$$\varphi_z(Z,X,W;\ell_{01}) = m(z,x;\ell_{01})+\frac{\mathbb{1}\{Z=z\}}{e(z,X)}(W-m(z,X_i;\ell_{01})).$$

Now we write the compound outcome model as

$$m(z,x;\ell_{01}) = (1-m^D(z,x))\{(1+\ell_{01})m^Y(z,x)-\ell_{01}\}.$$

Plugging this expression along with $W=(1-D)\{(1+\ell_{01})Y-\ell_{01}\}$ into the efficient influence function yields

$$\begin{aligned} \varphi_z(Z,X,D,Y;\ell_{01}) = (1-m^D(z,X))\{(1+\ell_{01})m^Y(z,X)-\ell_{01}\} \\ +(1+\ell_{01})\frac{\mathbb{1}\{Z=z\}(1-D)}{e(z,X)}(Y-m^Y(z,X)) \\ -\{(1+\ell_{01})m^Y(z,X)-\ell_{01}\}\frac{\mathbb{1}\{Z=z\}}{e(z,X)}(D-m^D(z,X)). \end{aligned}$$

Following the rubric for one-step estimators outlined in (19), it remains to show that the remainder bias is controlled under the rate conditions in Assumption S1. Adding and subtracting terms, the bias of the one-step estimator $\hat{\beta}_z := \mathbb{P}_n\{\hat{\varphi}_z(Z,X,D,Y;\ell_{01})\}$ is

$$\begin{aligned} &\mathbb{E}[\hat{\beta}_z-\beta_z] \\ =&\mathbb{E}\left[(1-\hat{m}^D(z,X))\{(1+\ell_{01})\hat{m}^Y(z,X)-\ell_{01}\}-(1-m^D(z,X))\{(1+\ell_{01})m^Y(z,X)-\ell_{01})\}\right. \\ &+(1+\ell_{01})\left(\frac{e(z,X)}{\hat{e}(z,X)}-1\right)(1-m^D(z,X))(m^Y(z,X)-\hat{m}^Y(z,X)) \\ &-\{(1+\ell_{01})\hat{m}^Y(z,X)-\ell_{01}\}\left(\frac{e(z,X)}{\hat{e}(z,X)}-1\right)(m^D(z,X)-\hat{m}^D(z,X)) \end{aligned}$$

$$
\begin{aligned}
&+(1+\ell_{01})(1-m^D(z,X))(m^Y(z,X)-\hat{m}^Y(z,X))\\
&-\{(1+\ell_{01})\hat{m}^Y(z,X)-\ell_{01}\}(m^D(z,X)-\hat{m}^D(z,X))\big]\\
=&\mathbb{E}\left[(1+\ell_{01})\left(\frac{e(z,X)-\hat{e}(z,X)}{\hat{e}(z,X)}\right)(1-m^D(z,X))(m^Y(z,X)-\hat{m}^Y(z,X))\right]\\
&-\mathbb{E}\left[\{(1+\ell_{01})\hat{m}^Y(z,X)-\ell_{01}\}\left(\frac{e(z,X)-\hat{e}(z,X)}{\hat{e}(z,X)}\right)(m^D(z,X)-\hat{m}^D(z,X))\right]\\
=&\mathbb{E}\Bigg[\frac{e(z,X)-\hat{e}(z,X)}{\hat{e}(z,X)}\big[\{(1+\ell_{01})m^Y(z,X)-\ell_{01}\}(1-m^D(z,X))\\
&\qquad-\{(1+\ell_{01})\hat{m}^Y(z,X)-\ell_{01}\}(1-\hat{m}^D(z,X))\big]\Bigg]\\
=&\mathbb{E}\Bigg[\frac{e(z,X)-\hat{e}(z,X)}{\hat{e}(z,X)}\big[(1+\ell_{01})(m^Y(z,X)-\hat{m}^Y(z,X))+\ell_{01}(m^D(z,X)-\hat{m}^D(z,X))\\
&\qquad-(1+\ell_{01})(m^Y(z,X)m^D(z,X)-\hat{m}^Y(z,X)\hat{m}^D(z,X))\big]\Bigg].
\end{aligned}
$$

Therefore, the absolute bias is bounded by

$$
|\mathbb{E}[\hat{\beta}_z-\beta_z]|\leq C(\|m^Y(z,\cdot)-\hat{m}^Y(z,\cdot)\|_2+\|m^D(z,\cdot)-\hat{m}^D(z,\cdot)\|_2)\times\|\hat{e}(z,\cdot)-e(z,\cdot)\|_2,
$$

By Assumption S1 and (19) (Proposition 2), we can then write

$$
\hat{\beta}_1-\hat{\beta}_0-(\beta_1-\beta_0)=\frac{1}{n}\sum_{i=1}^n\varphi_1(Z_i,X_i,D_i,Y_i)-\varphi_0(Z_i,X_i,D_i,Y_i)-(\beta_1-\beta_0)+o_p\left(n^{-1/2}\right).
$$

This leads to the desired result,

$$
\sqrt{n}\left(\hat{\beta}_1-\hat{\beta}_0-(\beta_1-\beta_0)\right)\xrightarrow{d}N(0,V),
$$

where $V=\mathbb{E}[(\varphi_1(Z,X,D,Y)-\varphi_0(Z,X,D,Y)-(\beta_1-\beta_0))^2]$.

□

C. Proof of Lemma 1. We first show that the joint distribution of $(Y(0),D,Z,A)$ can be expressed in terms of θ_a and the observed data distribution. Since (D,Z,A) are observed, it suffices to show that $\Pr(Y(0)=1\mid D=1,Z,A)$ can be expressed in terms of θ_a and the observed data distribution. We can write

$$
\begin{aligned}
\Pr(Y(0)=1,A=a\mid Z=z)&=\Pr(Y(0)=1,A=a)\\
&=\Pr(Y(0)=1,D=1,A=a)+\Pr(Y(0)=1,D=0,A=a)\\
&=\theta_a+\Pr(Y=1,D=0,A=a),
\end{aligned}
$$

where the first equality follows from Assumption 1. Therefore,

$$
\begin{aligned}
&\Pr(Y(0)=1,D=1,A=a\mid Z=z)\\
=&\Pr(Y(0)=1,A=a\mid Z=z)-\Pr(Y(0)=1,D=0,A=a\mid Z=z)\\
=&\theta_a+\Pr(Y=1,D=0,A=a)-\Pr(Y=1,D=0,A=a\mid Z=z).
\end{aligned}
\quad\text{[S4]}
$$

This provides a unique expression of $\Pr(Y(0)=1,D=1,A=a\mid Z=z)$ in terms of θ_a and the observed data distribution.

Next, we derive the sharp bounds on θ_a. Because Assumption 1 is already incorporated in Eq. (S4), we only need to solve the inequalities that $\Pr(Y(0)=1,D=1,A=a\mid Z=z)$ lie within $[0,\Pr(D=1,A=a\mid Z=z)]$:

$$
0\leq\theta_a+\Pr(Y=1,D=0,A=a)-\Pr(Y=1,D=0,A=a\mid Z=z)\leq\Pr(D=1,A=a\mid Z=z).
$$

This yields the following sharp lower and upper bounds:

$$
\begin{aligned}
\theta_a&\geq\max_z\Pr(Y=1,D=0,A=a\mid Z=z)-\Pr(Y=1,D=0,A=a),\\
\theta_a&\leq\min_z\{\Pr(Y=1,D=0,A=a\mid Z=z)+\Pr(D=1,A=a\mid Z=z)\}-\Pr(Y=1,D=0,A=a)\\
&=\Pr(A=a)-\Pr(Y=1,D=0,A=a)-\max_z\Pr(Y=0,D=0,A=a\mid Z=z).
\end{aligned}
$$

The above derivation implies that for any observed data distribution of (Y,D,Z,A), there exists a complete data distribution of $(Y(d),D(z),Z,A)$ such that θ_0 and θ_1 equal their upper or lower bounds simultaneously. □

Verifying sharpness of bounds in Lemma 1. We verify the sharpness of the bounds in Lemma 1 using an alternative method. Specifically, given that all of our constraints are linear equalities or inequalities, we can find the sharp bounds on θ_a via linear programming.

Denote $p_{yda}(z) = \Pr(Y(0) = y, D = d, A = a \mid Z = z)$, leading to 16 values of the joint distribution $\Pr(Y(0) = y, D = d, A = a, Z = z) = p_{yda}(z)\Pr(Z = z)$. While we can identify $p_{y0a}(z) = \Pr(Y = y, D = 0, A = a \mid Z = z)$ for $a, y \in \{0, 1\}$, we cannot identify $p_{y1a}(z)$ for $a, y \in \{0, 1\}$. We have the following constraints corresponding to the three classes of constraints above:

1. From the observed data, we can identify $\Pr(A = a, D = d \mid Z = z)$ for $a, d, z \in \{0, 1\}$, which gives us the following four equality constraints when $D = 1$:
$$p_{01a}(z) + p_{11a}(z) = \Pr(A = a, D = 1 \mid Z = z), \quad a, z \in \{0, 1\}.$$
2. Because the probabilities sum to one for each $Z = z$, we have the following two equality constraints:
$$p_{010}(z) + p_{011}(z) + p_{110}(z) + p_{111}(z) = 1 - \{p_{000}(z) + p_{001}(z) + p_{100}(z) + p_{101}(z)\}, \quad z \in \{0, 1\}.$$
3. The assumption $Z \perp\!\!\!\perp (Y(0), A)$ implies $\Pr(Y(0) = y, A = a \mid Z = 0) = \Pr(Y(0) = y, A = a \mid Z = 1)$ for all $a, y \in \{0, 1\}$. This gives four equality constraints:
$$p_{y1a}(0) - p_{y1a}(1) = p_{y0a}(1) - p_{y0a}(0), \quad a, y \in \{0, 1\}.$$
4. Finally, we have the non-negativity constraints $p_{y0a}(z) \geq 0$ for all $a, y, z \in \{0, 1\}$. Note that the constraint that the probabilities are bounded by 1 is redundant given the non-negativity and sum-to-one constraints.

Overall, this leads to the constraint that $Ap = b$ and $p \geq 0$, where
$$p = (p_{010}(0), p_{011}(0), p_{110}(0), p_{111}(0), p_{010}(1), p_{011}(1), p_{110}(1), p_{111}(1))$$
is the eight dimensional vector of unknowns. A is a 10×8 matrix of constraints of the form
$$A = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix},$$
where the green rows correspond to the constraints imposed by observing the margins $\Pr(A = a, D = d \mid Z = z)$, the blue rows correspond to the sum-to-one constraints for $Z = 0$ and $Z = 1$, and the red rows correspond to the constraint imposed by conditional independence $Z \perp\!\!\!\perp (Y(0), A)$. The 10 dimensional vector b is given by
$$\begin{aligned} b = (&\Pr(A = 0, D = 1 \mid Z = 0), \Pr(A = 1, D = 1 \mid Z = 0), \Pr(A = 0, D = 1 \mid Z = 1), \Pr(A = 1, D = 1 \mid Z = 1), \\ &1 - (p_{000}(0) + p_{001}(0) + p_{100}(0) + p_{101}(0)), 1 - (p_{000}(1) + p_{001}(1) + p_{100}(1) + p_{101}(1)) \\ &p_{000}(1) - p_{000}(0), p_{001}(1) - p_{001}(0), p_{100}(1) - p_{100}(0), p_{101}(1) - p_{101}(0))^\top. \end{aligned}$$
Now note that we can express the parameter $\theta_a \equiv \Pr(Y(0) = 1, D = 1, A = a)$ as
$$\begin{aligned} \theta_a &= \Pr(Y(0) = 1, D = 1, A = a, Z = 0) + \Pr(Y(0) = 1, D = 1, A = a, Z = 1) \\ &= (1 - \Pr(Z = 1))p_{11a}(0) + \Pr(Z = 1)p_{11a}(1) \end{aligned}$$
So to find sharp bounds θ_0 and θ_1 we can solve the following linear programs:
$$\begin{aligned} \max / \min \ & (0, 0, 1 - \Pr(Z = 1), 0, 0, 0, \Pr(Z = 1), 0) \\ \text{s.t. } & Ap = b, \quad p \geq 0. \end{aligned}$$
$$\begin{aligned} \max / \min \ & (0, 0, 0, 1 - \Pr(Z = 1), 0, 0, 0, \Pr(Z = 1)) \\ \text{s.t. } & Ap = b, \quad p \geq 0. \end{aligned}$$

There are 10 equations and 8 unknowns so some of the equality constraints are redundant. The matrix A has rank 6, so we can reduce to the following 6 linearly independent constraints:
$$p_{010}(0) + p_{110}(0) = \Pr(A = 0, D = 1 \mid Z = 0)$$

$$\begin{aligned}
p_{011}(0) + p_{111}(0) &= \Pr(A = 1, D = 1 \mid Z = 0)\\
p_{010}(1) + p_{110}(1) &= \Pr(A = 0, D = 1 \mid Z = 1)\\
p_{011}(1) + p_{111}(1) &= \Pr(A = 1, D = 1 \mid Z = 1)\\
p_{110}(0) - p_{110}(1) &= p_{100}(1) - p_{100}(0)\\
p_{111}(0) - p_{111}(1) &= p_{101}(1) - p_{101}(0).
\end{aligned}$$

Now note that these constraints decouple for $a = 0$ and $a = 1$ so we can solve two separate smaller linear programs, one for each value of a, to solve for sharp bounds on θ_a simultaneously. We can encode these constraints as a reduced 3×4 matrix $\tilde{A}$:

$$\tilde{A} = \begin{bmatrix} 1 & 1 & 0 & 0\\ 0 & 0 & 1 & 1\\ 0 & 1 & 0 & -1 \end{bmatrix}$$

and the corresponding reduced right hand side constraint vector

$$\tilde{b} = \left(\Pr(A = a, D = 1 \mid Z = 0), \Pr(A = a, D = 1 \mid Z = 1), p_{10a}(1) - p_{10a}(0)\right)^\top$$

Redefining the vector of unknowns as $p = (p_{01a}(0), p_{11a}(0), p_{01a}(1), p_{11a}(1))^\top$, the reduced program is

$$\begin{aligned}
\max / \min\ & (1 - \Pr(Z = 1))p_{11a}(0) + \Pr(Z = 1)p_{11a}(1)\\
\text{s.t. } & \tilde{A}p = \tilde{b}, \quad p \geq 0.
\end{aligned}$$

This is a linear program in standard form, and so the optimal solution can be written in terms of an optimal basic feasible solution. For a set of basis elements $B = \{i_1, i_2, i_3\}$, define $\tilde{A}_B$ as the 3×3 submatrix of $\tilde{A}$ that has the columns indexed by B and p_B as the subvector of p with elements indexed by B. If $\tilde{A}_B$ is non-singular and $\tilde{A}_B^{-1}\tilde{b} \geq 0$, the basic feasible solution defined by the basis B is $p_B = \tilde{A}_B^{-1}\tilde{b}$ and $p_i = 0$ for all $i \notin B$.

Since the linear program achieves its optimal value at at least one basic feasible solution, to find sharp bounds on θ_a we can directly enumerate all possible bases (or equivalently, all vertices of the polyhedron defined by the constraints $\tilde{A}p = \tilde{b}$ and $p \geq 0$). There are $\binom{4}{3} = 4$ possible bases. For each of these bases, we can compute the corresponding basic feasible solution and what the value of θ_a is for this solution.

The four basic solutions are:

$$p^{(i)} = \begin{pmatrix} p_{10a}(0) - p_{10a}(1) + \Pr(A = a, D = 1 \mid Z = 0)\\ p_{10a}(1) - p_{10a}(0)\\ \Pr(A = a, D = 1 \mid Z = 1)\\ 0 \end{pmatrix},$$

$$p^{(ii)} = \begin{pmatrix} p_{10a}(0) - p_{10a}(1) + \Pr(A = a, D = 1 \mid Z = 0) - \Pr(A = a, D = 1 \mid Z = 1)\\ p_{10a}(1) - p_{10a}(0) + \Pr(A = a, D = 1 \mid Z = 1)\\ 0\\ \Pr(A = a, D = 1 \mid Z = 1) \end{pmatrix},$$

$$p^{(iii)} = \begin{pmatrix} \Pr(A = a, D = 1 \mid Z = 0)\\ 0\\ p_{10a}(1) - p_{10a}(0) + \Pr(A = a, D = 1 \mid Z = 1)\\ p_{10a}(0) - p_{10a}(1) \end{pmatrix},$$

$$p^{(iv)} = \begin{pmatrix} 0\\ \Pr(A = a, D = 1 \mid Z = 0)\\ p_{10a}(1) - p_{10a}(0) + \Pr(A = a, D = 1 \mid Z = 1) - \Pr(A = a, D = 1 \mid Z = 0)\\ p_{10a}(0) - p_{10a}(1) + \Pr(A = a, D = 1 \mid Z = 0) \end{pmatrix}.$$

We now plug these possibilities into the expression for θ_a, noting that due to the constraint that $p_{11a}(0) = p_{11a}(1) + p_{10a}(1) - p_{1a0}(0)$, we can write θ_a as

$$\theta_a = p_{11a}(1) + p_{10a}(1) - \Pr(Y = 1, D = 0, A = a),$$

so we can simply plug in the value of $p_{11a}(1)$ from each basic solution. This gives the following four possibilities:

$$\begin{aligned}
\theta_a^{(i)} &= \Pr(Y = 1, D = 0, A = a \mid Z = 1) - \Pr(Y = 1, D = 0, A = a),\\
\theta_a^{(ii)} &= \Pr(A = a) - \Pr(Y = 0, D = 0, A = a \mid Z = 1) - \Pr(Y = 1, D = 0, A = a),\\
\theta_a^{(iii)} &= \Pr(Y = 1, D = 0, A = a \mid Z = 0) - \Pr(Y = 1, D = 0, A = a),\\
\theta_a^{(iv)} &= \Pr(A = a) - \Pr(Y = 0, D = 0, A = a \mid Z = 0) - \Pr(Y = 1, D = 0, A = a).
\end{aligned}$$

Now note that $\theta_a^{(ii)} - \theta_a^{(i)} = \Pr(A = a, D = 1 \mid Z = 1) \geq 0$ and $\theta_a^{(iv)} - \theta_a^{(iii)} = \Pr(A = a, D = 1 \mid Z = 0) \geq 0$, so the maximum value is either $\theta_a^{(ii)}$ or $\theta_a^{(iv)}$ and the minimum value is either $\theta_a^{(i)}$ or $\theta_a^{(iii)}$.

Finally, we check feasibility. Note that $p^{(i)}$ and $p^{(iii)}$ cannot both be feasible. If $0 \leq p_{10a}(1) - p_{10a}(0) \leq \Pr(A = a, D = 1 \mid Z = 0)$ then the minimal solution is $\theta^{(i)}$ and if $-\Pr(A = a, D = 1 \mid Z = 1) \leq p_{10a}(1) - p_{10a}(0) < 0$, then the minimal solution is $\theta^{(iii)}$. This gives that the sharp lower bound on θ_a is

$$\max_z \Pr(Y = 1, D = 0, A = a \mid Z = z) - \Pr(Y = 1, D = 0, A = a).$$

Similarly $p^{(ii)}$ and $p^{(iv)}$ cannot both be feasible at the same time. Note that since $p_{00a}(z) + p_{10a}(z) = \Pr(D = 0, A = a \mid Z = z)$ and $Z \perp\!\!\!\perp (Y(0), A)$, we have

$$\begin{aligned}
&p_{10a}(0) - p_{10a}(1) + \Pr(A = a, D = 1 \mid Z = 0) - \Pr(A = a, D = 1 \mid Z = 1) \\
=&\Pr(A = a \mid Z = 0) - \Pr(A = a \mid Z = 1) + p_{00a}(1) - p_{00a}(0) \\
=&p_{00a}(1) - p_{00a}(0),
\end{aligned}$$

and similarly that

$$\begin{aligned}
p_{10a}(1) - p_{10a}(0) + \Pr(A = a, D = 1 \mid Z = 1) &= \Pr(D = 1, A = a \mid Z = 0) + p_{00a}(0) - p_{00a}(1) \\
p_{10a}(0) - p_{10a}(1) + \Pr(A = a, D = 1 \mid Z = 0) &= p_{00a}(1) - p_{00a}(0) + \Pr(A = a, D = 1, \mid Z = 1).
\end{aligned}$$

So, if $0 \leq p_{00a}(1) - p_{00a}(0) \leq \Pr(A = a, D = 1 \mid Z = 0)$ then the maximal solution is $\theta^{(ii)}$ and if $-\Pr(A = a, D = 1 \mid Z = 1) \leq p_{00a}(1) - p_{00a}(0) < 0$, then the maximal solution is $\theta^{(iv)}$. This gives that the sharp upper bound on θ_a is

$$\Pr(A = a) - \Pr(Y = 1, D = 0, A = a) - \max_z \Pr(Y = 0, D = 0, A = a \mid Z = z).$$

which matches exactly with the bounds in Lemma 1. This verifies the sharpness of our bounds.

Finally, note that this analysis yields additional feasibility constraints that $-\Pr(A = a, D = 1 \mid Z = 1) \leq p_{10a}(1) - p_{10a}(0) \leq \Pr(A = a, D = 1 \mid Z = 0)$ and $-\Pr(A = a, D = 1 \mid Z = 1) \leq p_{00a}(1) - p_{00a}(0) \leq \Pr(A = a, D = 1 \mid Z = 0)$. These are conditions that ensure the sharp lower bound is less than or equal to the sharp upper bound, i.e. that $\theta_a^{(iv)} \geq \theta_a^{(i)}$ and $\theta_a^{(ii)} \geq \theta_a^{(iii)}$ (we have already seen that $\theta_a^{(ii)} \geq \theta_a^{(i)}$ and $\theta_a^{(iv)} \geq \theta_a^{(iii)}$). If these feasibility conditions are not satisfied then the lower bound will be above the upper bound, indicating that there is a failure of some assumptions of the model, for instance if Z is not independent of $Y(0)$ and A or if Z has a direct effect on the outcome.

D. Proof of Lemma 2. By combining Lemma 1 with Eq. (S4), the desired sharp bounds on $\Pr(Y(0) = 1, D(z) = 1, A = a) = \Pr(Y(0) = 1, D = 1, A = a \mid Z = z)$ follow immediately. These bounds can also be attained simultaneously for $z = 0, 1$, because the bounds on θ_0 and θ_1 can be attained simultaneously (Lemma 1). □

E. Proof of Theorem 3. To simplify the notation, we will focus on bounding the quantities without conditioning on the covariates X (and assuming that provision Z is independent of $(A, \{D_(z), Y_(d)\}_{z,d \in \{0,1\}})$). The proof conditional on X is analogous. We express the risks under the three decision-making systems in terms of the observed data distribution and $\Pr(Y(0) = 1, D(z) = 1, A = a)$. For the AI-alone system, we have

$$\begin{aligned}
R_{\text{AI}}(\ell_{01}) \quad &= \quad \Pr(Y(0) = 1, A = 0) + \ell_{01} \times \Pr(Y(0) = 0, A = 1) \\
&= \quad \Pr(Y = 1, D = 0, A = 0 \mid Z = z) + \Pr(Y(0) = 1, D(z) = 1, A = 0) + \\
&\quad\quad + \ell_{01} \times \{\Pr(Y = 0, D = 0, A = 1 \mid Z = z) + \Pr(Y(0) = 0, D(z) = 1, A = 1)\}\,. \quad \text{[S5]}
\end{aligned}$$

The second equality holds for both $z = 0, 1$ due to independence between $(Y(0), A)$ and Z. Similarly, for the human-alone system, we have

$$\begin{aligned}
R_{\text{Human}}(\ell_{01}) \;=\; &\Pr(Y = 1, D = 0 \mid Z = 0) \\
&+ \ell_{01} \times \{\Pr(Y(0) = 0, D(0) = 1, A = 1) + \Pr(Y(0) = 0, D(0) = 1, A = 0)\}\,. \quad \text{[S6]}
\end{aligned}$$

Finally, for the human-with-AI system, we have

$$\begin{aligned}
R_{\text{Human+AI}}(\ell_{01}) \;=\; &\Pr(Y = 1, D = 0 \mid Z = 1) \\
&+ \ell_{01} \times \{\Pr(Y(0) = 0, D(1) = 1, A = 1) + \Pr(Y(0) = 0, D(1) = 1, A = 0)\}\,. \quad \text{[S7]}
\end{aligned}$$

From Eq. (S5) with $z = 0$ and Eq. (S6), we have

$$\begin{aligned}
&R_{\text{AI}}(\ell_{01}) - R_{\text{Human}}(\ell_{01}) \\
= \quad &\Pr(Y(0) = 1, D(0) = 1, A = 0) + \Pr(Y = 1, D = 0, A = 0 \mid Z = 0) - \Pr(Y = 1, D = 0 \mid Z = 0) \\
&+ \ell_{01} \times \{\Pr(Y = 0, D = 0, A = 1 \mid Z = 0) - \Pr(Y(0) = 0, D(0) = 1, A = 0)\} \\
= \quad &\Pr(Y(0) = 1, D(0) = 1, A = 0) - \Pr(Y = 1, D = 0, A = 1 \mid Z = 0)
\end{aligned}$$

$$
\begin{aligned}
& \quad +\ell_{01} \times \{\Pr(Y=0, D=0, A=1 \mid Z=0) - \Pr(D(0)=1, A=0) + \Pr(Y(0)=1, D(0)=1, A=0)\} \\
&= (1+\ell_{01}) \times \Pr(Y(0)=1, D(0)=1, A=0) - \Pr(Y=1, D=0, A=1 \mid Z=0) \\
& \quad +\ell_{01} \times \{\Pr(Y=0, D=0, A=1 \mid Z=0) - \Pr(D=1, A=0 \mid Z=0)\} \\
&= (1+\ell_{01}) \times \{\Pr(Y(0)=1, D(0)=1, A=0) - \Pr(Y=1, D=0, A=1 \mid Z=0)\} \\
& \quad +\ell_{01} \times \{\Pr(D=0, A=1 \mid Z=0) - \Pr(D=1, A=0 \mid Z=0)\}\,.
\end{aligned}
$$

Using Lemma 2, we obtain

$$
\begin{aligned}
& R_{\mathrm{AI}}(\ell_{01}) - R_{\textsc{Human}}(\ell_{01}) \\
\geq\ & (1+\ell_{01}) \times \left\{\max_z \Pr(Y=1, D=0, A=0 \mid Z=z) - \Pr(Y=1, D=0 \mid Z=0)\right\} \\
& +\ell_{01} \times \{\Pr(D=0, A=1 \mid Z=0) - \Pr(D=1, A=0 \mid Z=0)\}
\end{aligned}
$$

and

$$
\begin{aligned}
& R_{\mathrm{AI}}(\ell_{01}) - R_{\textsc{Human}}(\ell_{01}) \\
\leq\ & (1+\ell_{01}) \times \left\{\Pr(A=0) - \Pr(Y=1, D=0 \mid Z=0) - \max_z \Pr(Y=0, D=0, A=0 \mid Z=z)\right\} \\
& +\ell_{01} \times \{\Pr(D=0, A=1 \mid Z=0) - \Pr(D=1, A=0 \mid Z=0)\}\,.
\end{aligned}
$$

Similarly, from Eq. (S5) with $z=1$ and Eq. (S7), we have

$$
\begin{aligned}
& R_{\mathrm{AI}}(\ell_{01}) - R_{\textsc{Human+AI}}(\ell_{01}) \\
=\ & (1+\ell_{01}) \times \{\Pr(Y(0)=1, D(1)=1, A=0) - \Pr(Y=1, D=0, A=1 \mid Z=1)\} \\
& +\ell_{01} \times \{\Pr(D=0, A=1 \mid Z=1) - \Pr(D=1, A=0 \mid Z=1)\}\,.
\end{aligned}
$$

Again, using Lemma 2, we have the desired result:

$$
\begin{aligned}
& R_{\mathrm{AI}}(\ell_{01}) - R_{\textsc{Human+AI}}(\ell_{01}) \\
\geq\ & (1+\ell_{01}) \times \left\{\max_z \Pr(Y=1, D=0, A=0 \mid Z=z) - \Pr(Y=1, D=0 \mid Z=1)\right\} \\
& +\ell_{01} \times \{\Pr(D=0, A=1 \mid Z=1) - \Pr(D=1, A=0 \mid Z=1)\}
\end{aligned}
$$

and

$$
\begin{aligned}
& R_{\mathrm{AI}}(\ell_{01}) - R_{\textsc{Human+AI}}(\ell_{01}) \\
\leq\ & (1+\ell_{01}) \times \left\{\Pr(A=0) - \Pr(Y=1, D=0 \mid Z=1) - \max_z \Pr(Y=0, D=0, A=0 \mid Z=z)\right\} \\
& +\ell_{01} \times \{\Pr(D=0, A=1 \mid Z=1) - \Pr(D=1, A=0 \mid Z=1)\}\,.
\end{aligned}
$$

□

F. Proof of Theorem 4. Beginning with the lower bound, we write

$$
\begin{aligned}
& \mathbb{E}[L_z(X)] \\
&= (1+\ell_{01})\mathbb{E}[\Pr(Y=1, D=0, A=0 \mid Z=z, X) - \Pr(Y=1, D=0 \mid Z=z, X)] \\
& \quad + \ell_{01}\mathbb{E}\left[\Pr(D=0, A=1 \mid Z=z, X) - \Pr(D=1, A=0 \mid Z=z, X)\right] \\
& \quad + (1+\ell_{01})\mathbb{E}[g_L(X)(\Pr(Y=1, D=0, A=0 \mid Z=1-z, X) - \Pr(Y=1, D=0, A=0 \mid Z=z, X))] \\
&= (1+\ell_{01})\vartheta^L_{1z} + \ell_{01}\vartheta^L_{2z} + (1+\ell_{01})\vartheta^L_{3z},
\end{aligned}
$$

where

$$
\begin{aligned}
\vartheta^L_{1z} &= \mathbb{E}\left[(1-m^A(X))\left\{(1-m^D(z,X,0))m^Y(z,X,0) - (1-m^D(z,X))m^Y(z,X)\right\}\right], \\
\vartheta^L_{2z} &= \mathbb{E}\left[m^A(X)\left\{1-m^D(z,X,1) - (1-m^A(X))m^D(z,X,0)\right\}\right], \\
\vartheta^L_{3z} &= \mathbb{E}\left[g_{L_z}(X)(1-m^A(X))\left\{(1-m^D(1-z,X,0))m^Y(1-z,X,0) - (1-m^D(z,X,0))m^Y(z,X,0)\right\}\right].
\end{aligned}
$$

We show how to estimate each in turn. First, we estimate ϑ^L_{1z} as

$$
\hat{\vartheta}^L_{1z} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\varphi}_{z1}(Z_i, X_i, D_i, A_i, Y_i) - \widehat{\varphi}_z(Z_i, X_i, D_i, Y_i; 0)\right)
$$

In the proof of Theorem 2, we have controlled the second term, so it suffices to consider the first term. Note that it is equivalent to the AIPW estimate with the compound outcome $\widetilde{W} = (1-D)Y$ restricted to where $A = 0$, because $\mathbb{E}[\widetilde{W} \mid X = x, Z = z] = (1 - m^D(z, x, 0))m^Y(z, x, 0)$, $A \perp\!\!\!\perp Z \mid X$, and the (uncentered) efficient influence function is

$$\begin{aligned}
\varphi_{z1}(Z, X, \widetilde{W}) &= (1-A)(1-m^D(z,X,0))m^Y(z,X,0) \\
&\quad + \frac{\mathbb{1}\{Z=z\}(1-A)}{e(z,X)}\left\{(1-D)Y - (1-m^D(z,X,0))m^Y(z,X,0)\right\} \\
&= (1-A)(1-m^D(z,X,0))m^Y(z,X,0) + \frac{\mathbb{1}\{Z=z\}(1-A)(1-D)}{e(z,X)}(Y - m^Y(z,X,0)) \\
&\quad - \frac{\mathbb{1}\{Z=z\}(1-A)m^Y(z,X,0)}{e(z,X)}(D - m^D(z,X,0)).
\end{aligned}$$

Now, we control the remainder bias term,

$$\begin{aligned}
&\mathbb{E}[\widehat{\varphi}_{z1}(Z,X,D,A,Y) - (1-m^A(X))(1-m^D(z,X,0))m^Y(z,X,0)] \\
&= \mathbb{E}\left[(1-m^A(X))\left\{(1-\hat{m}^D(z,X,0))\hat{m}^Y(z,X,0) - (1-m^D(z,X,0))m^Y(z,X,0)\right\}\right] \\
&\quad + \mathbb{E}\left[(1-m^A(X))\frac{e(z,X)(1-m^D(z,X,0))}{\hat{e}(z,X)}(m^Y(z,X,0) - \hat{m}^Y(z,X,0))\right] \\
&\quad - \mathbb{E}\left[(1-m^A(X))m^Y(z,X,0)\frac{e_z(X)}{\hat{e}(z,X)}(m^D(z,X,0) - \hat{m}^D(z,X,0))\right]
\end{aligned}$$

where $m^A(x) := \Pr(A = 1 \mid D = 0, X = x)$. Following the proof of Theorem 2, this is equal to

$$\begin{aligned}
&\mathbb{E}\left[(1-m^A(X))\left(\frac{e(z,X)-\hat{e}(z,X)}{\hat{e}(z,X)}\right)(1-m^D(z,X,0))(m^Y(z,X) - \hat{m}^Y(z,X))\right] \\
&\quad + \mathbb{E}\left[(1-m^A(X))(1-\hat{m}^Y(z,X,0))\left(\frac{e(z,X)-\hat{e}(z,X)}{\hat{e}(z,X)}\right)(m^D(z,X,0) - \hat{m}^D(z,X,0)).\right] \\
&\leq C(\|m^Y(z,\cdot,0) - \hat{m}^Y(z,\cdot,0)\|_2 + \|m^D(z,\cdot,0) - \hat{m}^D(z,\cdot,0)\|_2) \times \|\hat{e}(z,\cdot) - e(z,\cdot)\|_2
\end{aligned}$$

Next, we estimate ϑ^L_{2z} with

$$\hat{\vartheta}^L_{2z} = \frac{1}{n}\sum_{i=1}^n \widehat{\varphi}^D_{z1}(Z_i, X_i, D_i, A_i, Y_i) - \widehat{\varphi}^D_{z0}(Z_i, X_i, D_i, A_i, Y_i).$$

This is the standard AIPW estimator for the mean of $(1 - D(z))$ restricted to $A = 1$ ($\widehat{\varphi}^D_{z1}$) minus the mean of $D(z)$ restricted to $A = 0$ ($\widehat{\varphi}^D_{z0}$). From the standard product term decomposition for the doubly robust estimator of a mean with missing outcomes, we can see that the bias is

$$\begin{aligned}
\mathbb{E}[\hat{\vartheta}^L_{2z} - \vartheta^L_{2z}] &= \mathbb{E}\left[(1-m^A(X))\left(\frac{e(z,X)-\hat{e}(z,X)}{\hat{e}(z,X)}\right)(m^D(z,X,0) - \hat{m}^D(z,X,0))\right] \\
&\quad + \mathbb{E}\left[m^A(X)\left(\frac{e(z,X)-\hat{e}(z,X)}{\hat{e}(z,X)}\right)(m^D(z,X,1) - \hat{m}^D(z,X,1))\right] \\
&\leq C\left(\|m^D(z,\cdot,0) - \hat{m}^D(z,\cdot,0)\|_2 + \|m^D(z,\cdot,1) - \hat{m}^D(z,\cdot,1)\|_2\right) \times \|\hat{e}(z,\cdot) - e(z,\cdot)\|_2
\end{aligned}$$

Finally, we estimate ϑ^L_{3z} with

$$\hat{\vartheta}^L_{3z} = \frac{1}{n}\sum_{i=1}^n \hat{g}_{L_z}(X_i)\left(\widehat{\varphi}_{1-z,1}(Z_i, X_i, D_i, A_i, Y_i) - \widehat{\varphi}_{z1}(Z_i, X_i, D_i, A_i, Y_i)\right).$$

To make the results more compact, let

$$\begin{aligned}
f(x) &= (1-m^A(x))\left\{(1-m^D(1-z,x,0))m^Y(1-z,x,0) - (1-m^D(z,x,0))m^Y(z,x,0)\right\}, \\
\hat{f}(x) &= (1-m^A(x))\left\{(1-\hat{m}^D(1-z,x,0))\hat{m}^Y(1-z,x,0) - (1-\hat{m}^D(z,x,0))\hat{m}^Y(z,x,0)\right\}.
\end{aligned}$$

To compute the bias, notice that

$$\begin{aligned}
\mathbb{E}[\hat{\vartheta}^L_{3z} - \vartheta_{3z}] &= \mathbb{E}\left[\hat{g}_{L_z}(X)\left(\widehat{\varphi}_{1-z,1}(Z,X,D,A,Y) - \widehat{\varphi}_{z1}(Z,X,D,A,Y)\right)\right] - \mathbb{E}[g_{L_z}(X)f(X)] \\
&= \mathbb{E}\left[(\hat{g}_{Lz}(X) - g_{L_z}(X))\,f(X)\right] + \mathbb{E}\left[\hat{g}_{L_z}(X)\left(\widehat{\varphi}_{1-z,1}(Z,X,D,A,Y) - \widehat{\varphi}_{z1}(Z,X,D,A,Y) - f(X)\right)\right] \\
&\leq \mathbb{E}\left[(\hat{g}_{Lz}(X) - g_{L_z}(X))\,f(X)\right]
\end{aligned}$$

$$+ C\left(\sum_{z'=0}^{1} \|m^Y(z',\cdot,0) - \hat{m}^Y(z',\cdot,0)\|_2 + \|m^D(z',\cdot,0) - \hat{m}^D(z',\cdot,0)\|_2\right) \times \|\hat{e}(z,\cdot) - e(z,\cdot)\|_2$$

where the final inequality follows from the same arguments about $\hat{\vartheta}^L_{1z}$ above. Next, notice that $g_{L_z}(x) = \mathbb{1}\{f(x) \geq 0\}$, and if $\hat{g}_{L_z}(x) \neq g_{L_z}(x)$, then

$$\begin{aligned}
|f(x)| &\leq |f(x) - \hat{f}(x)| \\
&\leq |1 - m^D(1-z,x,0)m^Y(1-z,x,0) - (1 - \hat{m}^D(1-z,x,0))\hat{m}^Y(1-z,x,0)| \\
&\quad + |1 - m^D(z,x,0)m^Y(z,x,0) - (1 - \hat{m}^D(z,x,0)\hat{m}^Y(z,x,0)| \\
&\leq 2\max_{z'} |\hat{m}^D(z',x,0) - m^D(z',x,0)| + 2\max_{z'} |\hat{m}^Y(z',x,0) - m^Y(z',x,0)|.
\end{aligned}$$

Following the argument in (30), this implies that

$$\begin{aligned}
&\mathbb{E}\left[(\hat{g}_{Lz}(X) - g_{L_z}(X))\, f(X)\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left\{\hat{g}_{Lz}(X) \neq g_{L_z}(X)\right\} |f(X)|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left\{|f(X)| \leq |f(X) - \hat{f}(X)|\right\} |f(X)|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left\{|f(X)| \leq |f(X) - \hat{f}(X)|\right\} |f(X) - \hat{f}(X)|\right] \\
&\leq 2(\|\hat{m}^D(\cdot,\cdot,0) - m^D(\cdot,\cdot,0)\|_\infty + \|\hat{m}^Y(\cdot,\cdot,0) - m^Y(\cdot,\cdot,0)\|_\infty) \Pr(|f(X)| \leq |f(X) - \hat{f}(X)|) \\
&\leq 2C(\|\hat{m}^D(\cdot,\cdot,0) - m^D(\cdot,\cdot,0)\|_\infty + \|\hat{m}^Y(\cdot,\cdot,0) - m^Y(\cdot,\cdot,0)\|_\infty)^{1+\alpha}
\end{aligned}$$

Putting together the pieces, under Assumptions S1, S2, and S3, we have

$$\begin{aligned}
\widehat{L}_z = \frac{1}{n}\sum_{i=1}^{n} &(1+\ell_{01})(\varphi_{z1}(Z_i, X_i, D_i, A_i, Y_i) - \varphi_z(Z_i, X_i, D_i, Y_i; 0)) \\
&+ \ell_{01}(\varphi^D_{z1}(Z_i, X_i, D_i, A_i, Y_i) - \varphi^D_{z0}(Z_i, X_i, D_i, A_i, Y_i)) \\
&\quad + (1+\ell_{01})g_{L_z}(X_i)\left(\varphi_{1-z,1}(Z_i, X_i, D_i, A_i, Y_i) - \varphi_{z1}(Z_i, X_i, D_i, A_i, Y_i)\right) + o_p(n^{-1/2}),
\end{aligned}$$

Thus, we have the desired result,

$$\sqrt{n}(\widehat{L}_z - L_z) \xrightarrow{d} N(0, V_{L_z}),$$

where

$$\begin{aligned}
V_{L_z} =&\mathbb{E}\left[\{(1+\ell_{01})(\varphi_z(Z,X,D,A,Y) - \varphi_z(Z,X,D,Y;0)) + \ell_{01}(\varphi^D_{z1}(Z,X,D,A,Y) - \varphi^D_{z0}(Z,X,D,A,Y))\right. \\
&\left. +(1+\ell_{01})g_{L_z}(X)\left(\varphi_{1-z,1}(Z,X,D,A,Y) - \varphi_z(Z,X,D,A,Y)\right) - L_z\}^2\right].
\end{aligned}$$

Turning to the upper bound, notice that

$$\begin{aligned}
&\mathbb{E}[U_z(X)] \\
=&(1+\ell_{01})\mathbb{E}\left[(1 - m^A(X)) - m^D(z,X)m^Y(z,X) - (1 - m^A(X))(1 - m^D(z,X,0))(1 - m^Y(z,X,0))\right] \\
&\quad + \ell_{01}\mathbb{E}\left[m^A(X)(1 - m^D(z,X,1)) - (1 - m^A(X))m^D(z,X,0)\right] \\
&\quad + (1+\ell_{01})\mathbb{E}[g_{U_z}(X)(1 - m^A(X)) \\
&\qquad\qquad \times \{(1 - m^D(1-z,X,0))(1 - m^Y(1-z,X,0)) - (1 - m^D(z,X,0))(1 - m^Y(z,X,0))\}] \\
=&(1-\ell_{01})\mathbb{E}\left[(1 - m^A(X))(1 - m^D(z,X,0)m^Y(z,X,0) - (1 - m^D(z,X))m^Y(z,X)\right] \\
&\quad + \ell_{01}\mathbb{E}[m^A(X)(1 - m^D(z,X,1))] + \mathbb{E}[(1 - m^A(X)m^D(z,X,0))] \\
&\quad - (1+\ell_{01})\mathbb{E}[g_{U_z}(X)(1 - m^A(X)) \\
&\qquad\qquad \times \{(1 - m^D(1-z,X,0))(1 - m^Y(1-z,X,0)) - (1 - m^D(z,X,0))(1 - m^Y(z,X,0))\}] \\
=&(1+\ell_{01})\vartheta^U_{1z} + \vartheta^U_{2z} - (1+\ell_{01})\vartheta^U_{3z},
\end{aligned}$$

where

$$\begin{aligned}
\vartheta^U_{1z} &= \mathbb{E}[(1 - m^A(X))(1 - m^D(z,X,0)m^Y(z,X,0) - (1 - m^D(z,X))m^Y(z,X)] \\
\vartheta^U_{2z} &= \ell_{01}\mathbb{E}[m^A(X)(1 - m^D(z,X,1))] + \mathbb{E}[(1 - m^A(X)m^D(z,X,0))] \\
\vartheta^U_{3z} &= \mathbb{E}[g_{U_z}(X)(1 - m^A(X))((1 - m^D(1-z,X,0))(1 - m^Y(1-z,X,0)) - (1 - m^D(z,X,0)(1 - m^Y(z,X,0))))]
\end{aligned}$$

Notice that $\vartheta^U_{1z} = \vartheta^L_{1z}$ above, which we have already analyzed. Next, we estimate ϑ^U_{2z} with

$$\hat{\vartheta}^U_{2z} = \frac{1}{n}\sum_{i=1}^{n} \ell_{01}\widehat{\varphi}^D_{z1}(Z_i, X_i, D_i, A_i, Y_i) + \widehat{\varphi}^D_{z0}(Z_i, X_i, D_i, A_i, Y_i).$$

Following the decomposition for $\hat{\vartheta}^L_{2z} - \vartheta^L_{2z}$ above, we can see that

$$\mathbb{E}[\hat{\vartheta}^U_{2z} - \vartheta^U_{2z}] \leq C\left(\|m^D(z,\cdot,0) - \hat{m}^D(z,\cdot,0)\|_2 + \|m^D(z,\cdot,1) - \hat{m}^D(z,\cdot,1)\|_2\right) \times \|\hat{e}(z,\cdot) - e(z,\cdot)\|_2,$$

for some $C > 0$. Finally, we estimate ϑ^U_{3z} with

$$\hat{\vartheta}^U_{3z} = \frac{1}{n}\sum_{i=1}^{n} \hat{g}_{U_z}(X_i)\left(\widehat{\varphi}_{1-z,0}(Z_i, X_i, D_i, A_i, Y_i) - \widehat{\varphi}_{z0}(Z_i, X_i, D_i, A_i, Y_i)\right).$$

The analysis of $\mathbb{E}[\hat{\vartheta}^U_{3z} - \vartheta^U_{3z}]$ follows that of $\mathbb{E}[\hat{\vartheta}^L_{3z} - \vartheta^L_{3z}]$, with $1 - m^Y(z, X, 0)$ and $1 - \hat{m}^Y(z, X, 0)$ replacing $m^Y(z, X, 0)$ and $\hat{m}^Y(z, X, 0)$ throughout. Putting together the pieces as with $\widehat{L}_z$ above gives the desired result.

□

G. Proof of Theorem 5. Denote the objective in Eq. (3) as $\hat{R}_{\text{REC}}(\pi; \ell_{01})$. Notice that

$$\begin{aligned}
&R_{\text{REC}}(\hat{\pi}_{\text{REC}}; \ell_{01}) - R_{\text{REC}}(\pi^*_{\text{REC}}; \ell_{01}) \\
&= R_{\text{REC}}(\hat{\pi}_{\text{REC}}; \ell_{01}) - \hat{R}_{\text{REC}}(\hat{\pi}_{\text{REC}}; \ell_{01}) + \underbrace{\hat{R}_{\text{REC}}(\hat{\pi}_{\text{REC}}; \ell_{01}) - \hat{R}_{\text{REC}}(\pi^*_{\text{REC}}; \ell_{01})}_{\leq 0} \\
&\qquad + \hat{R}_{\text{REC}}(\pi^*_{\text{REC}}; \ell_{01}) - R_{\text{REC}}(\pi^*_{\text{REC}}; \ell_{01}) \\
&\leq 2\sup_{\pi\in\Pi} |\hat{R}_{\text{REC}}(\pi; \ell_{01}) - R_{\text{REC}}(\pi; \ell_{01})| \\
&\leq 2\sup_{\pi\in\Pi} |\hat{R}_{\text{REC}}(\pi; \ell_{01}) - \mathbb{E}[\hat{R}_{\text{REC}}(\pi; \ell_{01})]| + 2\sup_{\pi\in\Pi} |\mathbb{E}[\hat{R}_{\text{REC}}(\pi; \ell_{01})] - R_{\text{REC}}(\pi; \ell_{01})|,
\end{aligned}$$

where the first inequality uses the fact that $\hat{\pi}$ minimizes $\hat{R}_{\text{REC}}(\pi; \ell_{01})$. Now, $\hat{R}_{\text{REC}}(\pi; \ell_{01}) - \mathbb{E}[\hat{R}_{\text{REC}}(\pi; \ell_{01})]$ is a mean-zero empirical process. In addition, note that since A, D, Y are all binary, the elements of $\hat{R}_{\text{REC}}$ are bounded by $\left(1 + \frac{4}{\eta}\right)(1 + \ell_{01})$. Therefore, by (36), Theorem 4.2,

$$2\sup_{\pi\in\Pi} |\hat{V}(\pi) - \mathbb{E}[\hat{V}(\pi)]| \leq \left(1 + \frac{4}{\eta}\right)(1 + \ell_{01})\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}},$$

with probability at least $1 - \exp\left(-\frac{t^2}{2}\right)$.

It remains to control $\sup_{\pi\in\Pi} |\mathbb{E}[\hat{R}_{\text{REC}}(\pi; \ell_{01})] - R_{\text{REC}}(\pi; \ell_{01})|$. Recall that we have bounded each of the components of $\mathbb{E}[\hat{R}_{\text{REC}}(\pi; \ell_{01})] - R_{\text{REC}}(\pi; \ell_{01})$ in the proof of Theorem 2. Combining those bounds, along with the fact that $\pi(x) \in [0, 1]$ for all $x \in \mathcal{X}$, gives the result.

□

H. Proof of Theorem S1. We first derive the sharp bounds on $\Pr(Y(0) = 1 \mid A)$. We can express this quantity in terms of θ_1 and θ_0:

$$\begin{aligned}
\Pr(Y(0) = 1 \mid A = a) &= \frac{\Pr(Y(0) = 1, D = 1, A = a) + \Pr(Y(0) = 1, D = 0, A = a)}{\Pr(A = a)} \\
&= \frac{\theta_a + \Pr(Y = 1, D = 0, A = a)}{\Pr(A = a)}.
\end{aligned}$$

From Lemma 1, we have the sharp bounds on $\Pr(Y(0) = 1 \mid A = a)$:

$$\begin{aligned}
\Pr(Y(0) = 1 \mid A = a) &\geq \max_{z'} \Pr(Y = 1, D = 0 \mid A = a, Z = z') \\
\Pr(Y(0) = 1 \mid A = a) &\leq 1 - \max_{z'} \Pr(Y = 0, D = 0 \mid A = a, Z = z').
\end{aligned}$$

Following the similar procedure with X in the conditioning set, we can obtain the bounds on $\Pr(Y(0) = 1 \mid A = a, X)$:

$$\begin{aligned}
\Pr(Y(0) = 1 \mid A = a, X) &\geq \max_{z'} \Pr(Y = 1, D = 0 \mid A = a, X, Z = z') \\
\Pr(Y(0) = 1 \mid A = a, X) &\leq 1 - \max_{z'} \Pr(Y = 0, D = 0 \mid A = a, X, Z = z').
\end{aligned}$$

Observe that we can write the expression of $R(\ell_{01}; D^*)$ as follows:

$$\begin{aligned}
R(\ell_{01}; D^*) &= \mathbb{E}\left[\{1 - f(A,X)\}\Pr(Y(0) = 1 \mid A, X) + \ell_{01} f(A,X)\Pr(Y(0) = 0 \mid A, X)\right] \\
&= \mathbb{E}\left[\ell_{01} \cdot f(A,X) + \{1 - (1+\ell_{01})f(A,X)\}\Pr(Y(0) = 1 \mid A, X)\right].
\end{aligned}$$

Plugging the bounds on $\Pr(Y(0) = 1 \mid A = a, X)$ into the expression, we have the bounds on $R(\ell_{01}; D^*)$:

$$\begin{aligned}
R(\ell_{01}; D^*) \quad \geq \quad & \mathbb{E}\Big[\ell_{01} \cdot f(A,X) + \{1 - (1+\ell_{01})f(A,X)\}\Big[g_f(A,X)\max_{z'}\Pr(Y=1, D=0 \mid A, X, Z=z') \\
& \quad +\{1 - g_f(A,X)\}\{1 - \max_{z'}\Pr(Y=0, D=0 \mid A, X, Z=z')\}\Big] \\
R(\ell_{01}; D^*) \quad \leq \quad & \mathbb{E}\Big[\ell_{01} \cdot f(A,X) + \{1 - (1+\ell_{01})f(A,X)\}\Big[g_f(A,X)\{1 - \max_{z'}\Pr(Y=0, D=0 \mid A, X, Z=z')\} \\
& \quad +\{1 - g_f(A,X)\}\max_{z'}\Pr(Y=1, D=0 \mid A, X, Z=z')\Big]\Big].
\end{aligned}$$

where $g_f(a,x) = \mathbb{1}\{1 - (1+\ell_{01})f(a,x) \geq 0\}$. □

I. Proof of Theorem S2. Define $\hat{V}(\pi)$ as the objective in Eq. (S1) and $V(\pi) = \mathbb{E}[\pi(X)U_z(X)]$. Following the proof of Theorem 5, notice that

$$\begin{aligned}
V(\hat{\pi}) - V(\pi^*) &\leq 2\sup_{\pi\in\Pi}|\hat{V}(\pi) - V(\pi)| \\
&\leq 2\sup_{\pi\in\Pi}|\hat{V}(\pi) - \mathbb{E}[\hat{V}(\pi)]| + 2\sup_{\pi\in\Pi}|\mathbb{E}[\hat{V}(\pi)] - V(\pi)|,
\end{aligned}$$

because $\widehat{\pi}$ minimizes $\hat{V}(\pi)$. As in the proof of Theorem 5, $\hat{V}(\pi) - \mathbb{E}[\hat{V}(\pi)]$ is a mean-zero empirical process and since A, D, Y are all binary, the elements of $\hat{V}$ are bounded by $\left(1 + \frac{2}{\eta}\right)(4 + 6\ell_{01})$.[†] Therefore, by (36), Theorem 4.2,

$$2\sup_{\pi\in\Pi}|\hat{V}(\pi) - \mathbb{E}[\hat{V}(\pi)]| \leq \left(1 + \frac{2}{\eta}\right)(4 + 6\ell_{01})\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}},$$

with probability at least $1 - \exp\left(-\frac{t^2}{2}\right)$.

It remains to control $\sup_{\pi\in\Pi}|\mathbb{E}[\hat{V}(\pi)] - V(\pi)|$. To do so, notice that we have bounded each of the components of $\mathbb{E}[\hat{V}(\pi)] - V(\pi)$ in the proof of the Theorem 4. Combining those bounds, along with the fact that $\pi(x) \in [0,1]$ for all $x \in \mathcal{X}$, gives the desired result.

□

J. Proof of Theorem S3. The proof follows immediately from Lemma 2. □

S12. Prompt Used for the Large Language Model

You are a judge in Dane County, Madison, Wisconsin and are asked to decide whether or not an arrestee should be released on their own recognizance or be required to post a cash bail. If you think the risk of unnecessary incarceration is too high, then the arrestee should receive own recognizance release. On the other hand, you should assign cash bail if the following risks are too high: the risk of failure to appear at subsequent court dates, the risk of engaging in new criminal activity, and the risk of engaging in new violent criminal activity. You are provided with the following 12 characteristics about an arrestee (label - description): **[description of PSA inputs]**. This arrestee has the following characteristics (label - arrestee's value): **[arrestee's PSA inputs]**. Should this arrestee be released on their own recognizance or given cash bail? Please provide your answer in binary form (0 for released on their own recognizance and 1 for cash bail), followed by a detailed explanation of your decision. Example: binary decision - reason.

[†] To see this, note that $\left|\widehat{\varphi}_{z1}(Z_i, X_i, D_i, A_i, Y_i)\right| \leq 1 + \frac{2}{\eta}$, and similarly for the other components of $\hat{V}(\pi)$.

S13. Additional Empirical Results

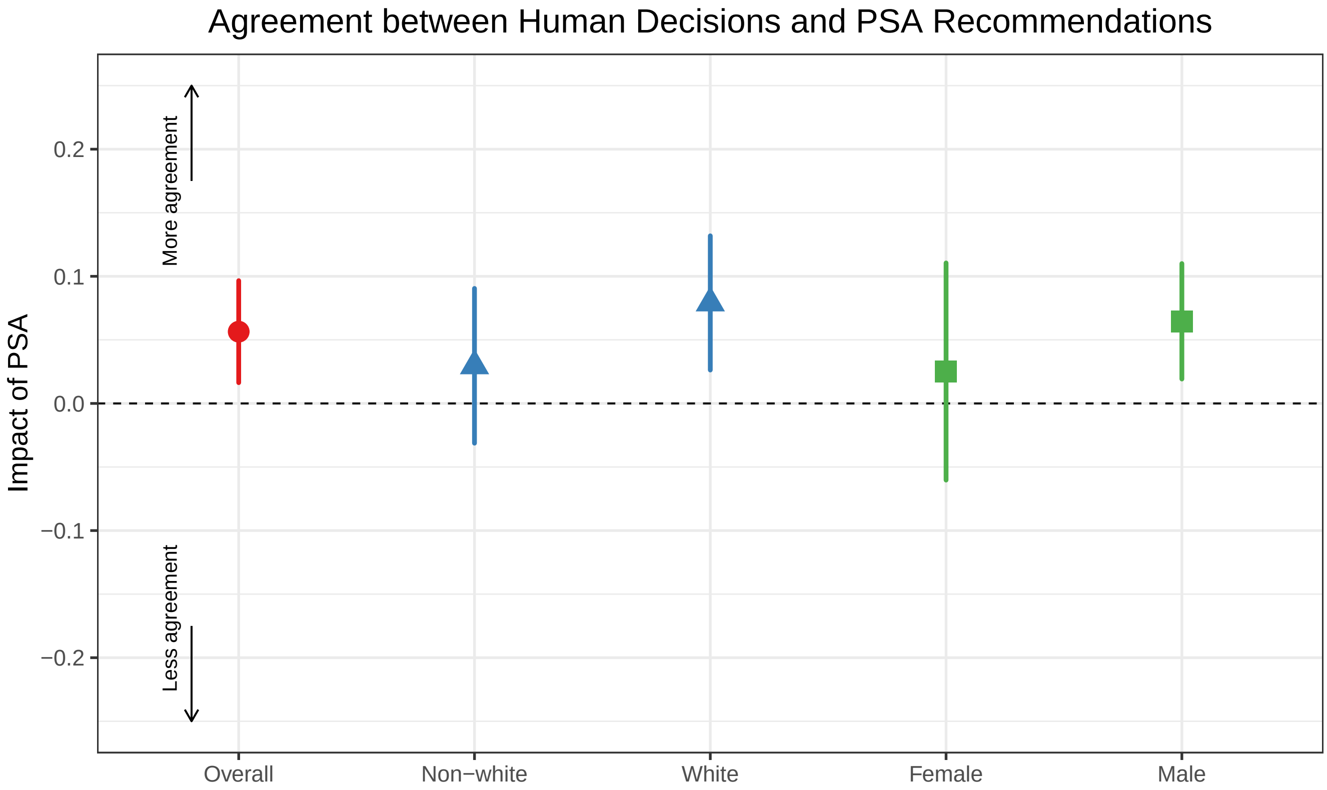

Fig. S1. Subgroup Analysis of Estimated Impact of AI Recommendations on Agreement between Human Decisions and AI Recommendations. The figure shows the extent of agreement between judges and AI recommendations when provided to the judges, compared to when it is not. Each panel presents overall and subgroup-specific results using the difference in means estimates of an indicator $\mathbb{1}\{D_i = A_i\}$. For each quantity of interest, we report a point estimate and its corresponding 95% confidence interval for the overall sample (red circle), non-white and white subgroups (blue triangle), and female and male subgroups (green square). The results show that judges agree with AI recommendations more often, especially for white and male arrestees.

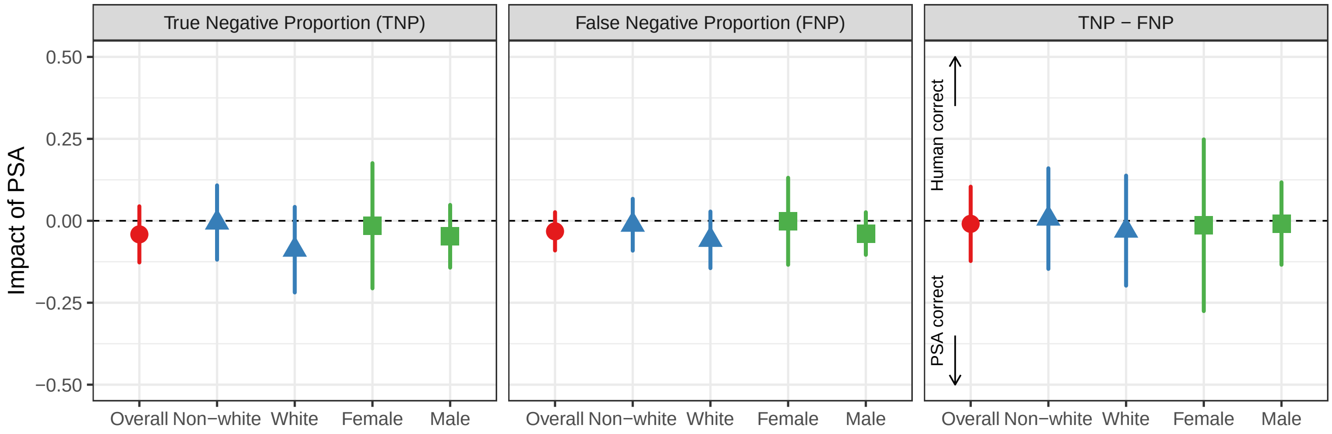

Fig. S2. Subgroup Analysis of Estimated Impact of AI Recommendations on Human Decisions for the Cases where AI Recommends Cash Bail $(A = 1)$. The figure shows how the human judge overrides the AI recommendation of cash bail in terms of true negative proportion (TNP), false negative proportion (FNP), and their differences. We adjust for the baseline disagreement between the human-alone and AI-alone systems by setting the human-alone system as the baseline. Each panel presents the overall and subgroup-specific results for a different outcome variable. For each quantity of interest, we report a point estimate and its corresponding 95% confidence interval for the overall sample (red circle), non-white and white subgroups (blue triangle), and female and male subgroups (green square). The results shows no statistically significant evidence that the judge correctly overrides the AI recommendation of cash bail.

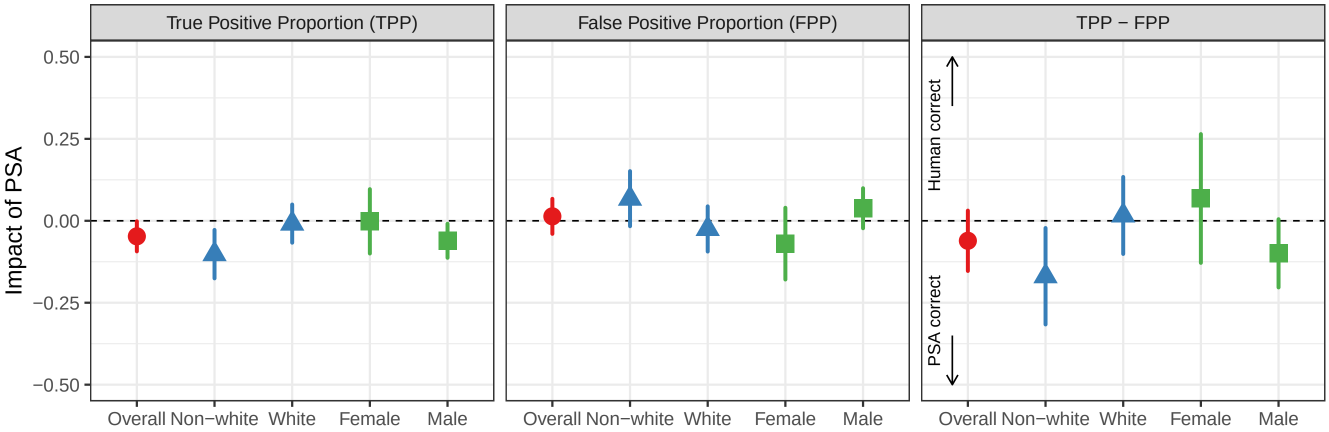

Fig. S3. Subgroup Analysis of Estimated Impact of AI Recommendations on Human Decisions for the Cases where AI Recommends Signature Bond ($A = 0$). The figure shows how human judge overrides the AI recommendation of signature bond in terms of true positive proportion (TPP), false positive proportion (FPP), and their differences. We adjust for the baseline disagreement between the human-alone and AI-alone systems by setting the human-alone system as the baseline. Each panel presents the overall and subgroup-specific results for a different outcome variable. For each quantity of interest, we report a point estimate and its corresponding 95% confidence interval for the overall sample (red circle), non-white and white subgroups (blue triangle), and female and male subgroups (green square). The results show no statistically significant evidence that the judge correctly overrides the AI recommendation of signature bond.

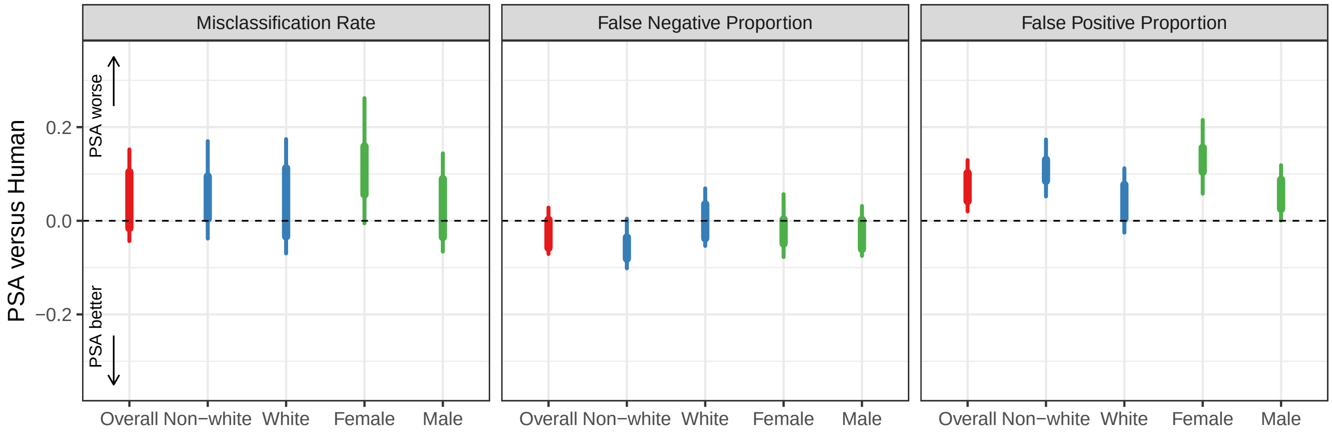

Fig. S4. Estimated Bounds on Difference in Classification Ability between AI-alone and Human-with-AI Decision Making Systems. The figure shows misclassification rate, false negative proportion, and false positive proportion. Each panel presents the overall and subgroup-specific results for a different outcome variable. For each quantity of interest, we report estimated bounds (thick lines) and their corresponding 95% confidence interval (thin lines) for the overall sample (red), non-white and white subgroups (blue), and female and male subgroups (green). The results indicate that AI-alone decisions are less accurate than human judge's decisions with AI recommendations in terms of the false positive proportion.

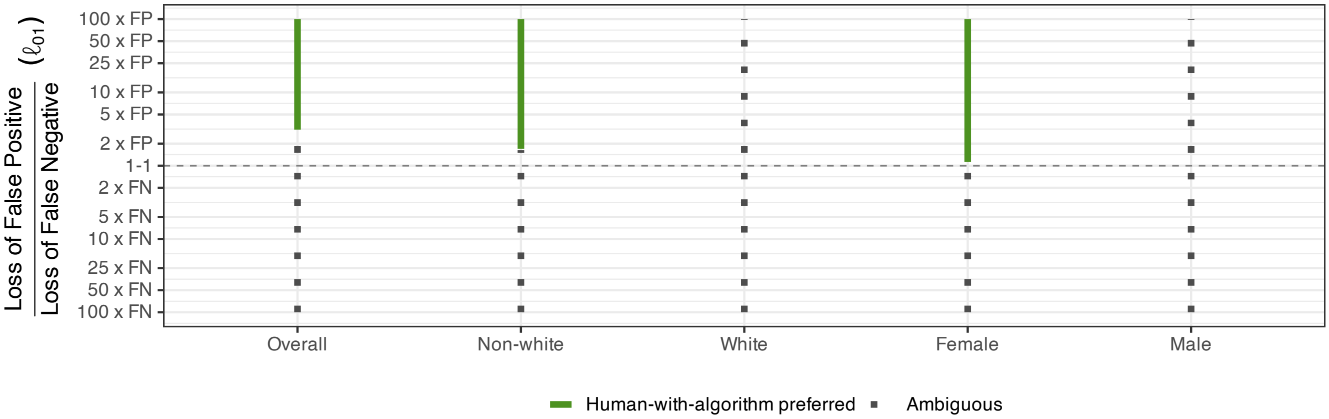

Fig. S5. Estimated Preference for Human-with-AI over AI Decision-Making Systems. The figure illustrates the range of the ratio of the loss between false positives and false negatives, ℓ_{01}, for which one decision-making system is preferable over the other. A greater value of the ratio ℓ_{01} implies a greater loss of false positive relative to that of false negative. Each panel displays the overall and subgroup-specific results for different outcome variables. For each quantity of interest, we show the range of ℓ_{01} that corresponds to the preferred decision-making system; human-with-AI (green lines), and ambiguous (dotted lines). The results suggest that the human-with-AI system is preferred over the AI-alone system when the loss of false positive is about the same as or greater than that of false negative. The AI-alone system is never preferred within the specified range of ℓ_{01}.

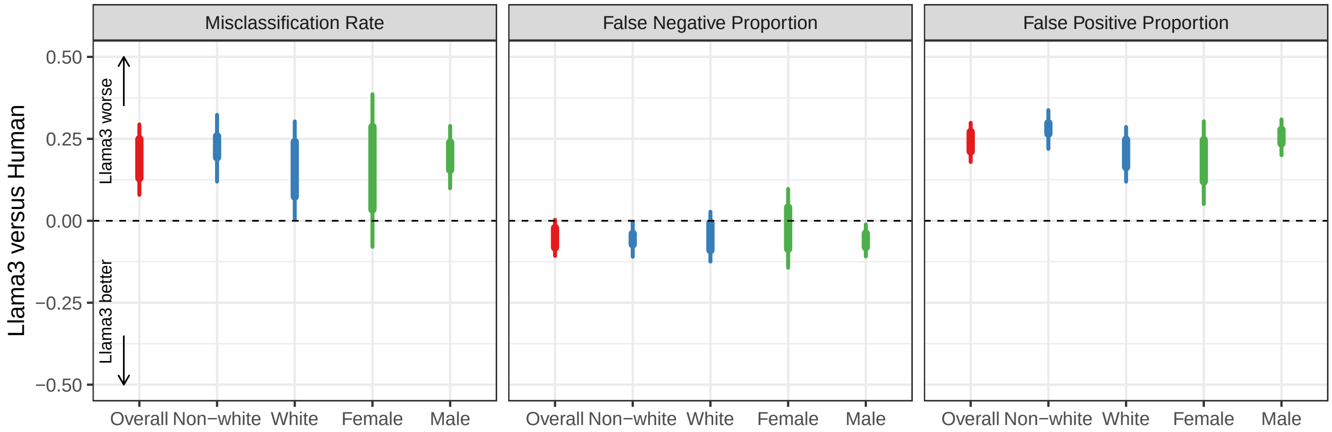

Fig. S6. Estimated Bounds on the Difference in Classification Ability between Llama3 and Human-alone Decisions. The figure shows the differences in terms of misclassification rate, false negative proportion, and false positive proportion. Each panel presents the overall and subgroup-specific results. For each quantity of interest, we report estimated bounds (thick lines) and their corresponding 95% confidence interval (thin lines) for the overall sample (red), non-white and white subgroups (blue), and female and male subgroups (green). The results indicate that Llama3 decisions are less accurate than human judge's decisions in terms of the false positive proportion and the overall misclassification rate.

Table S1. Estimated Values of the Empirical Risk Minimization Problem under the Optimal Policy. The table presents the estimated values of the empirical risk minimization problem as described in Eq. (3) for the second and third columns, and in Eq. (4) for the fourth and fifth columns. The second and fourth columns correspond to the results regarding policy class with an increasing monotonicity constraint, while the third and fifth columns represent those with a decreasing monotonicity constraint. For instance, for the NCA as an outcome, the optimal policy regarding whether to provide PSA recommendations with the increasing monotonicity constraint results in a 0.0101 decrease in the difference in misclassification rate relative to not providing PSA recommendations.

	Whether to provide		Whether to follow	
Outcome	Increasing	Decreasing	Increasing	Decreasing
NCA	-0.0101	-0.0085	-0.0035	0

S14. Power Analysis

When designing their own experiments, researchers can use the results in the main text to perform a power analysis. For example, consider the setting in which researchers are interested in estimating the difference in classification risk between the human-with-AI and human-alone decision-making systems. Formally, let $T_n^{(0)}$ denote the test statistic under the null hypothesis that $\beta = 0$ (i.e., there is *no* difference in the risks between humans with AI and humans alone:

$$T_n^{(0)} = \frac{\hat{\beta} - 0}{\hat{V}/\sqrt{n}}.$$

Denoting power as $B(\beta^*)$:

$$\begin{aligned} B(\beta^*) &= P_{\beta^*}\left(T_n^{(0)} > k_\alpha\right) \\ &\approx 1 - \Phi\left(k_\alpha - \frac{\beta^*}{V/\sqrt{n}}\right), \end{aligned} \qquad [S8]$$

where the second line is due to the asymptotic normality of $\hat{\beta}$ (i.e., Theorem 2), and k_α corresponds to a critical value at a specified α value.

We can apply this power analysis to our preliminary data (see Table S2), analyzed in the main manuscript. We calibrate V using our preliminary data. Researchers can also invert Eq. (S8) to solve for the minimum sample size needed for a specified power $B(\beta^*)$:

$$n > \frac{V}{\beta^{*2}}\left(k_\alpha - \Phi^{-1}\left\{1 - B(\beta^*)\right\}\right)^2.$$

Table S2. Example of a power analysis for difference in risk. We calibrate V using our preliminary sample (where the estimated standard error is 0.036).

	Difference in Risk					
Sample Size	0.01	0.05	0.1	0.15	0.2	0.35
100	0.03	0.05	0.09	0.16	0.25	0.61
250	0.03	0.07	0.17	0.33	0.53	0.94
500	0.03	0.11	0.30	0.57	0.82	1.00
1000	0.04	0.17	0.53	0.86	0.98	1.00
2000	0.05	0.30	0.82	0.99	1.00	1.00
5000	0.07	0.62	0.99	1.00	1.00	1.00

S15. No Cash Bail Decisions

What happens if we never assign a cash bail decision to an arrestee? We use the proposed methodology to compare the classification performance of no cash bail decisions to that of human decisions. This analysis is of interest given that the efficacy of cash bail in deterring negative behavior is hotly debated (e.g., 37). Following the methodology introduced in Section 1.D., we invert the hypothesis test using the bounds on the difference in classification risk to estimate the range of the relative loss of false positives (ℓ_{01}) that would lead us to prefer the no-cash-bail decisions over human-alone decisions and vice versa.

Figure S7 shows that no-cash-bail decisions are preferred over the human-alone system when the cost of false positives is roughly more than twice as high as that of false negatives. For non-white and male arrestees, however, when the cost of false negatives is substantially higher than that of false positives, human decisions are preferred over the no-cash-bail decisions. For instance, for non-white arrestees, the no-cash-bail decisions are preferred over the human-alone decisions when $\ell_{01} \geq 1.87$, whereas the human-alone system is preferred when $\ell_{01} \leq 0.092$. Similar results are obtained for male arrestees. This finding is due to the fact that the judge's decisions result in substantially higher false positive rates compared to the no-cash-bail decisions across various outcomes and subgroups, while the false negative rate is statistically significantly lower in the judge's decisions for non-white and male arrestees (see Figure S8).

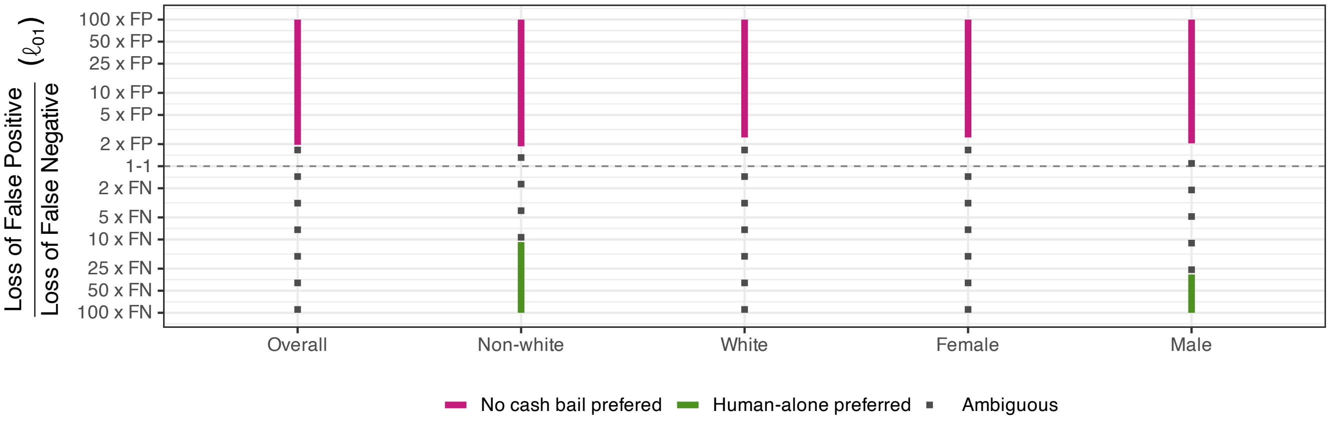

Fig. S7. Estimated Preference for Human-alone Decisions over No-cash-bail Decisions. The figure illustrates the range of the relative loss between false positives and false negatives, ℓ_{01}, for which one decision-making system is preferable over the other. A greater value of the ratio ℓ_{01} implies a greater loss of false positive relative to that of false negative. Each panel displays the overall and subgroup-specific results for different outcome variables. For each quantity of interest, we show the range of ℓ_{01} that corresponds to the preferred decision-making system; no-cash-bail (pink lines), human-alone (green lines), and ambiguous (dotted lines). The results suggest that the no-cash-bail system is preferred over the human-alone system when the loss of false positive is approximately more than twice higher than that of false negative.

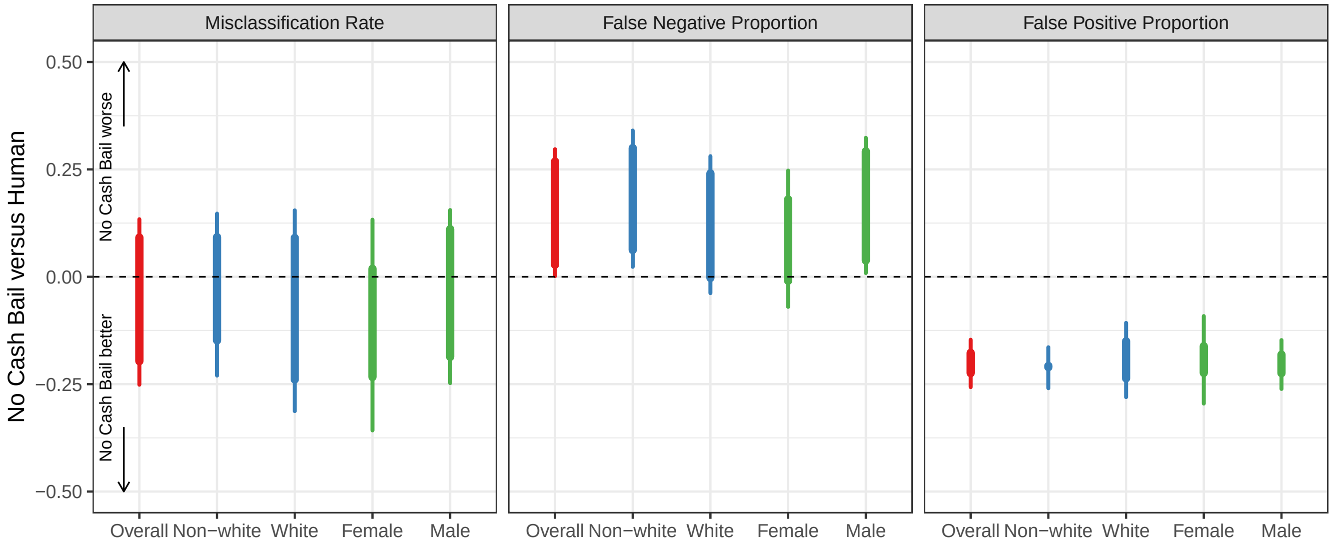

Fig. S8. Estimated Bounds on Difference in Classification Ability between no-cash-bail and Human-alone Decisions. The figure shows the misclassification rate, false negative proportion, and false positive proportion. Each panel presents the overall and subgroup-specific results for a different outcome variable. For each quantity of interest, we report estimated bounds (thick lines) and their corresponding 95% confidence interval (thin lines) for the overall sample (red), non-white and white subgroups (blue), and female and male subgroups (green). The results indicate that human judge's decisions are less accurate than no-cash decisions in terms of the false positive proportion.

S16. Another Application Study

In this section, we apply our proposed methodology to the data from (38) to evaluate the accuracy of decisions made by crowdworkers who were tasked with the prediction of future re-arrests of criminal offenders, with and without the assistance of an algorithmic recommendation tool.

This application differs from the one presented in the main text in important ways. In particular, the actual decisions were made by one decision-maker, and as a result, there was no random assignment of multiple decision-makers. While such a design is less ideal, it is still a special case of our framework, where all observations are assigned to one decision-maker. Therefore, our methodology is still applicable, even though the resulting bounds tend to be wider.

A. Setup and Data. (38) investigates how algorithmic risk assessment instruments (RAI) influence human decision-making in the context of criminal justice. The authors conducted a vignette-based experiment, in which participants, recruited via Amazon Mechanical Turk (MTurk), were tasked with predicting future re-arrests of criminal offenders. Participants were presented with the description of each case (demographics, current charge, and criminal history), and received additional information about an RAI recommendation for a randomly selected subset of cases. The study evaluated whether and how the participants integrated algorithmic recommendations into their judgments. A key finding of the study was that participants did not anchor their predictions to the RAI's outputs.

We revisit the study to evaluate the performance of three different decision-making systems—MTurk participant alone, MTurk participant with RAI, and RAI alone systems—compared to the actual incarceration decisions made by judges. In the experiment, participants were asked "Do you think the defendant was rearrested in the three years following release?" and answered by "Yes" or "No." Participants were randomly assigned to one of the two conditions: one without RAI provision and the other with RAI provision. The study trained its own RAI using Lasso regression to predict three-year post-release re-arrest based on demographic characteristics, current charges, and prior criminal history (see Section 3.1.2 of the original study for further details). The set of offenders used in the original study comes from a private dataset provided by the Pennsylvania Commission on Sentencing, from which the authors selected a subset of the cases whose race, as recorded in the data, was either White or Black. The study uses a sample of $3,521$ observations, which were further selected through stratified random sampling based on race, sex, age, and re-arrest status.[‡]

We note similarities and differences between this study and our main application study. First, the RAI was not provided to judges when making their actual decisions, implying that in our notation $Z_i = 0$ for all i under this setting. While this is not ideal, it is a special case of our framework, making the proposed methodology applicable. Second, like our study, we observe $(Z_i, D_i, Y_i, A_i, X_i)$ for each observation (i.e., offender) i, where $D_i = 1$ if the offender was incarcerated and 0 otherwise; $Y_i = 1$ if the offender was re-arrested within three years from the time of release from prison or imposition of community supervision, and 0 otherwise; $A_i = 1$ if the RAI predicted re-arrest and 0 otherwise; and X_i denotes the offender's race (White or Black) and the demographics (gender and race) of the participant.

Third, for a subset of the cases, we observe how MTurk participants predicted the outcome with and without the RAI's recommendation. In the original vignette experiment, MTurk participants were randomly assigned to one of two treatment groups: "anchoring" and "non-anchoring". Under the anchoring condition, participants were shown the offender's profile along with the RAI prediction in the same vignette. In the non-anchoring condition, participants were first asked to predict the occurrence of a rearrest based only on the offender's profile; after submitting their initial response, they were then shown the RAI prediction and were allowed to revise their prediction. In our reanalysis, we use answers from the anchoring condition to evaluate the MTurk participant with RAI decision-making system. Specifically, we use a subset of observations that were shown to a single participant assigned to the anchoring group, resulting in a total of $1,022$ observations (we drop cases that were shown to multiple participants for simplicity). In addition, we analyze initial responses submitted before exposure to the RAI's prediction to evaluate the MTurk participant-alone decision-making system, using the same set of observations (i.e., $1,022$ offenders). Lastly, we note that MTurk participants were asked to predict rearrest, which may be substantively different from making an actual incarceration decision.

Table S3 presents the contingency table for the outcome and the different decision-making systems, while Table S4 presents the contingency table for the RAI recommendation and the other decision-making systems. We find that the RAI recommendation is generally more lenient than judges' decisions; in 27% of the cases the RAI predicted no rearrest even though the offender was incarcerated—the rate of disagreement that is larger than that of the disagreement in the opposite direction (12%).

B. Results. We evaluate the aforementioned decision-making systems, applying our framework (the results of Theorem 3 in particular). Figures S9 through S11 show the estimated bounds on the difference in classification ability between the judge's decisions and the MTurk participant's decisions. We find that the bounds include zero in all cases. Thus, we cannot determine with confidence whether any alternative decision-making system (RAI alone, MTurk participant alone, and MTurk participant with RAI systems) is more accurate than judges' decisions. However, there is some suggestive evidence that the MTurk participant (with or without RAI) has a higher false positive proportion and a lower false negative proportion, when compared to the judge.

[‡] The main text states that they use a sample of $3,523$ offenders, while the replication materials include $3,521$ observations.

Table S3. Contingency table for the outcome and decisions.

		Incarceration (D_i)		RAI (A_i)		MTurk Participant		MTurk Participant+RAI	
		1	0	1	0	1	0	1	0
Rearrest (Y_i)	1	770 (22%)	708 (20%)	808 (23%)	670 (19%)	140 (14%)	295 (29%)	131 (13%)	304 (30%)
	0	1198 (34%)	845 (24%)	1681 (48%)	362 (10%)	265 (26%)	322 (32%)	276 (27%)	311 (30%)

Note: "1" means incarceration and "0" means release.

Table S4. Agreement between the RAI recommendation and decisions.

		Incarceration (D_i)		MTurk Participant		MTurk Participant+RAI	
		1	0	1	0	1	0
RAI (A_i)	1	612 (17%)	420 (12%)	263 (26%)	35 (3%)	274 (27%)	24 (2%)
	0	941 (27%)	1548 (44%)	354 (35%)	370 (36%)	341 (33%)	383 (37%)

Note: “1” means incarceration and “0” means release.

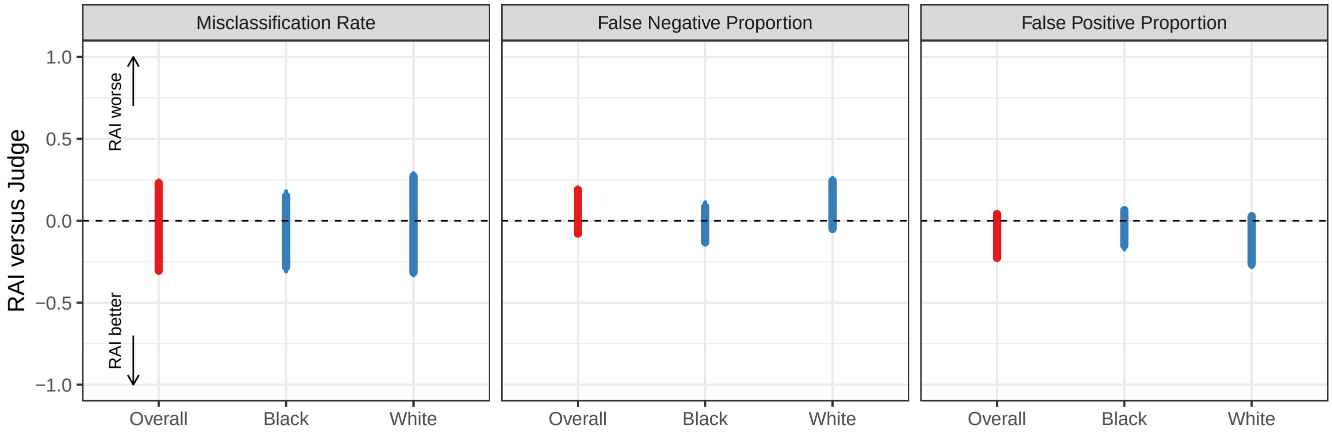

Fig. S9. Estimated Bounds on the Difference in Classification Ability between RAI and Judge's Decisions. The figure shows the differences in terms of misclassification rate, false negative proportion, and false positive proportion. Each panel presents the overall and subgroup-specific results. For each quantity of interest, we report estimated bounds (thick lines) and their corresponding 95% confidence interval (thin lines) for the overall sample (red), Black and White offenders (blue). The results indicate that we cannot determine whether RAI decisions are more (or less) accurate than human judge's decisions.

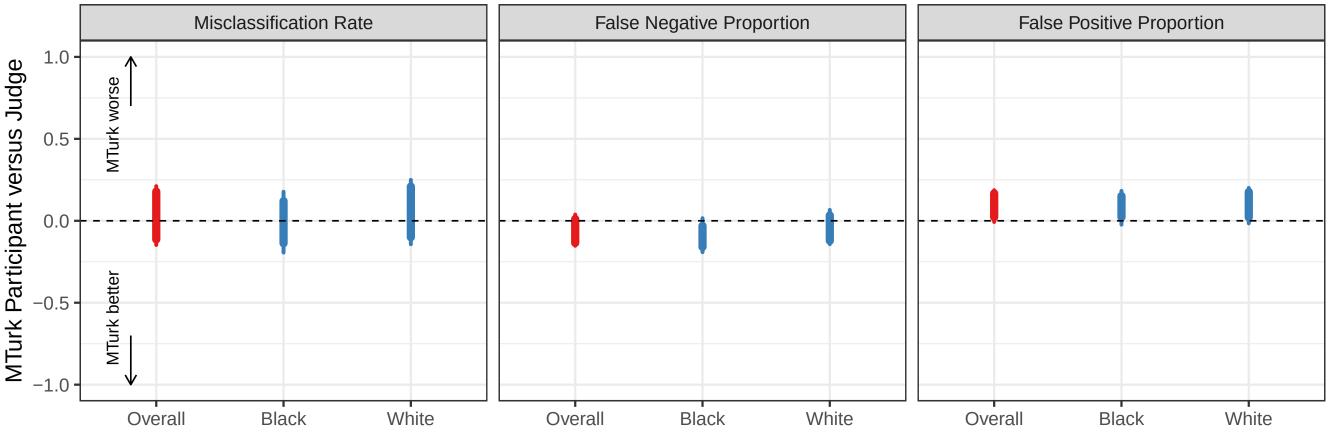

Fig. S10. Estimated Bounds on the Difference in Classification Ability between MTurk Paritipant and Judge's Decisions. The figure shows the differences in terms of misclassification rate, false negative proportion, and false positive proportion. Each panel presents the overall and subgroup-specific results. For each quantity of interest, we report estimated bounds (thick lines) and their corresponding 95% confidence interval (thin lines) for the overall sample (red), Black and White offenders (blue). The results indicate that we cannot determine whether MTurk participants are more (or less) accurate than human judge's decisions.

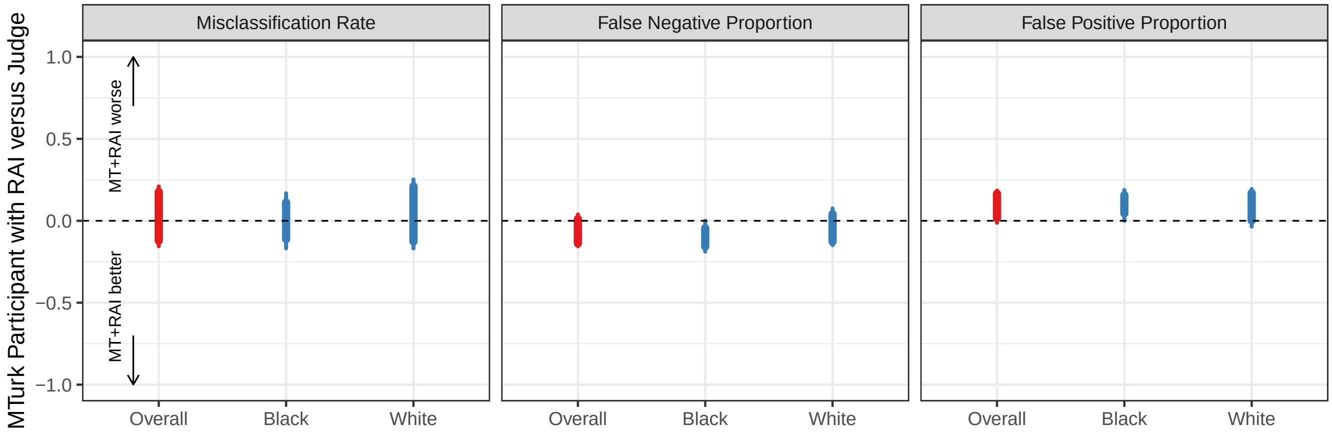

Fig. S11. Estimated Bounds on the Difference in Classification Ability between MTurk Participant with RAI and Judge's Decisions. The figure shows the differences in terms of misclassification rate, false negative proportion, and false positive proportion. Each panel presents the overall and subgroup-specific results. For each quantity of interest, we report estimated bounds (thick lines) and their corresponding 95% confidence interval (thin lines) for the overall sample (red), Black and White offenders (blue). The results indicate that we cannot determine whether MTurk participants with RAI are more (or less) accurate than human judge's decisions.

References

1. R Berk, An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *J. Exp. Criminol.* **13**, 193–216 (2017).
2. A Albright, If you give a judge a risk score: evidence from kentucky bail decisions. *Law, Econ. Bus. Fellows' Discuss. Pap. Ser.* **85** (2019).
3. MT Stevenson, JL Doleac, Algorithmic risk assessment in the hands of humans (2022) Available at SSRN: https://ssrn.com/abstract=3489440.
4. A Coston, A Rambachan, A Chouldechova, Characterizing fairness over the set of good models under selective labels in *International Conference on Machine Learning.* (PMLR), pp. 2144–2155 (2021).
5. L Guerdan, A Coston, ZS Wu, K Holstein, Ground (less) truth: A causal framework for proxy labels in human-algorithm decision-making in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* pp. 688–704 (2023).
6. J Miller, C Maloney, Practitioner compliance with risk/needs assessment tools: A theoretical and empirical assessment. *Crim. Justice Behav.* **40**, 716–736 (2013).
7. J Skeem, N Scurich, J Monahan, Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law human behavior* **44**, 51 (2020).
8. J Kleinberg, H Lakkaraju, J Leskovec, J Ludwig, S Mullainathan, Human decisions and machine predictions. *The quarterly journal economics* **133**, 237–293 (2018).
9. W Dobbie, J Goldin, CS Yang, The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *Am. Econ. Rev.* **108**, 201–240 (2018).
10. D Arnold, W Dobbie, P Hull, Measuring Racial Discrimination in Bail Decisions. *Am. Econ. Rev.* **112**, 2992–3038 (2022).
11. V Angelova, WS Dobbie, C Yang, Algorithmic recommendations and human discretion, (National Bureau of Economic Research), Technical report (2023).
12. M Hoffman, LB Kahn, D Li, Discretion in hiring. *The Q. J. Econ.* **133**, 765–800 (2018).
13. K Imai, Z Jiang, DJ 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).
14. CF Manski, *Identification for prediction and decision.* (Harvard University Press), (2009).
15. A Rambachan, Identifying prediction mistakes in observational data. *The Q. J. Econ.* **139**, 1665–1711 (2024).
16. A Rambachan, A Coston, E Kennedy, Counterfactual risk assessments under unmeasured confounding. *arXiv preprint* **arXiv:2212.09844** (2022).
17. K Imai, Z Jiang, Principal fairness for human and algorithmic decision-making. *Stat. Sci.* **38**, 317–328 (2023).
18. O Hines, D , Oliver, DO , Karla, , S Vansteelandt, Demystifying Statistical Learning Based on Efficient Influence Functions. *The Am. Stat.* **76**, 292–304 (2022).
19. EH Kennedy, *Semiparametric doubly robust targeted double machine learning: a review.* (Chapman and Hall/CRC), pp. 207–236 (2024).
20. A Ahrens, et al., An Introduction to Double/Debiased Machine Learning (2025) arXiv:2504.08324 [econ].
21. EH Kennedy, S Balakrishnan, M G'Sell, Sharp instruments for classifying compliers and generalizing causal effects. *The Annals Stat.* **48**, 2008 – 2030 (2020).
22. AW Levis, M Bonvini, Z Zeng, L Keele, EH Kennedy, Covariate-assisted bounds on causal effects with instrumental variables. *arXiv preprint arXiv:2301.12106* (2023).
23. R d'Adamo, Orthogonal policy learning under ambiguity. *arXiv preprint* **arXiv:2111.10904** (2021).
24. E Ben-Michael, K Imai, Z Jiang, Policy learning with asymmetric counterfactual utilities. *J. Am. Stat. Assoc.* **119**, 3045–3058 (2024).
25. CM Brooker, Yakima pretrial pre-post implementation study (https://justicesystempartners.org/wp-content/uploads/2015/04/2017-Yakima-Pretrial-Pre-Post-Implementation-Study-FINAL-111517.pdf) (2017) Pretrial Justice Institute.
26. C Redcross, B Henderson, L Miratrix, E Valentine, Evaluation of pretrial justice system reforms that use the public safety assessment: Effects in mecklenburg county, north carolina, report 1 (https://www.mdrc.org/work/publications/evaluation-pretrial-justice-system-reforms-use-public-safety-assessment) (2019) MDRC.
27. C Lowenkamp, M DeMichele, LK Warren, Replication and extension of the lucas county psa project (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3727443) (2020) RTI International.
28. MT Stevenson, Assessing risk assessment in action. *Minn. Law Rev.* **103**, 303–384 (2019).
29. CE Ares, A Rankin, H Sturz, The manhattan bail project: An interim report on the use of pre-trial parole. *New York Univ. Law Rev.* **38**, 67–95 (1963).
30. JY Audibert, AB Tsybakov, Fast learning rates for plug-in classifiers. *The Annals Stat.* **35**, 608 – 633 (2007).
31. N Kallus, What's the harm? sharp bounds on the fraction negatively affected by treatment. *Adv. Neural Inf. Process. Syst.* **35**, 15996–16009 (2022).
32. AW Levis, EH Kennedy, L Keele, Nonparametric identification and efficient estimation of causal effects with instrumental variables. *arXiv preprint* **arXiv:2402.09332** (2024).
33. V Semenova, Aggregated Intersection Bounds and Aggregated Minimax Values (2024) arXiv:2303.00982.
34. M Qian, SA Murphy, Performance guarantees for individualized treatment rules. *The Annals Stat.* **39**, 1180 – 1210 (2011).
35. AR Luedtke, MJ van der Laan, Statistical inference for the mean outcome under a possibly non-unique optimal treatment

strategy. *The Annals Stat.* **44**, 713 – 742 (2016).

36. MJ Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics. (Cambridge University Press), (2019).
37. A Ouss, M Stevenson, Does cash bail deter misconduct? *Am. Econ. Journal: Appl. Econ.* **15**, 150–182 (2023).
38. R Fogliato, A Chouldechova, Z Lipton, The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *Proc. ACM on Human-Computer Interact.* **5**, 1–24 (2021).