Evaluating AI and Machine Learning Algorithms for Causal Inference

Kosuke Imai

Harvard University

February 10, 2026

Distinguished Speaker Series, Statistical Horizons

Motivation

- Rise of causal machine learning (causal ML)
  1. heterogeneous treatment effects
  2. individualized treatment rules
- Rise of Artificial Intelligence (AI)
  1. recommendation systems
  2. automated decision-making systems
- A healthy skepticism:

  causal ML and AI may be biased and cause undesirable outcomes

- Need for statistical evaluation
  1. Did my ML algorithm accurately detect heterogeneous effects?
  2. How would my individualized treatment rule perform once deployed in the real world?
  3. Does AI help humans make better decisions?
  4. What happens if we completely outsource our decisions to AI?

Overview of Today's Seminar

1. Statistical evaluation of heterogeneous treatment effects discovered by causal ML
    - How well are heterogeneous treatment effects estimated?
    - How can I identify exceptional responders with statistical guarantee?
2. Statistical evaluation of individualized treatment rules derived by causal ML
    - How do ITRs perform in practice (relative to one another)?
    - What is the value of personalization?
3. Statistical evaluation of recommendation systems derived by AI and ML algorithms
    - Does AI help humans make better decisions?
    - When should humans ignore or follow AI recommendations?

We will focus on experimental studies though we will briefly discuss observational studies

Part I: Heterogeneous Treatment Effects

Heterogeneous Treatment Effects

- Motivation: the same treatment may affect different individuals differently
- Individual Treatment Effect (ITE):

$$\tau_i := Y_i(1) - Y_i(0)$$

  - binary treatment: $T_i \in \{0, 1\}$
  - potential outcomes: $Y_i(1)$ and $Y_i(0)$
  - ITE is unobservable: fundamental problem of causal inference

- Conditional Average Treatment Effect (CATE)

$$\tau(\boldsymbol{x}) \; := \; \mathbb{E}[Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}] \quad \text{where} \quad \boldsymbol{x} \in \mathcal{X}$$

  - characterizes how ITE varies as a function of $\boldsymbol{X}_i$ *on average*
  - what types of people are likely to benefit from and be harmed by the treatment?
  - *descriptive* (rather than causal) statements about ITE

- Assumption: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid \boldsymbol{X}_i = \boldsymbol{x}$ and $0 < \pi(\boldsymbol{x}) < 1$ for all $\boldsymbol{x}$

Subgroup Analysis and Pre-registration

- If we have a hypothesis about some group-specific effects:
  1. choose $\boldsymbol{X}$ based on this hypothesis
  2. stratify the data and estimate the ATE within each strata
  3. compare the ATE between groups
- Problem: multiple testing, "p-hacking"
- Solution: Pre-register hypotheses and analyses
  - standard in medicine, has become a norm in social sciences
  - repositories
    - Evidence in Governance and Politics (EGAP)
    - American Economic Association (AEA)
    - Registry for International Development Impact Evaluations (RIDIE)
- Pre-registration solves commitment and transparency problems
- It does not solve the statistical problem of multiple testing
- It still requires the use of statistical methods that control
  - FWER (family-wise error rate): probability of making *any* type I error
  - FDR (false discovery rate): expected proportion of type I error among all rejections

Machine Learning for Heterogeneous Causal Effects

- Motivation:
  1. avoid strong modeling assumptions $\rightsquigarrow$ data-driven approach
  2. avoid false discoveries $\rightsquigarrow$ avoid over-fitting via regularization
- Difference between standard prediction and heterogeneous treatment effect estimation
  - predict observed outcome: $\rightsquigarrow$ use $\boldsymbol{X}_i$ to predict Y_i
  - predict unobservable ITE: $\rightsquigarrow$ use $\boldsymbol{X}_i$ to predict $\tau_i := Y_i(1) - Y_i(0)$
- Is CATE a good predictor of ITE?

$$\begin{aligned}\text{Mean squared error} &= \mathbb{E}[(\tau_i - \hat{\tau}(\boldsymbol{x}))^2 \mid \boldsymbol{X}_i = \boldsymbol{x}] \\ &= \underbrace{\mathbb{E}[(\tau_i - \tau(\boldsymbol{x}))^2 \mid \boldsymbol{X}_i = \boldsymbol{x}]}_{\text{within-group heterogeneity}} + \underbrace{\mathbb{E}[(\tau(\boldsymbol{x}) - \hat{\tau}(\boldsymbol{x}))^2 \mid \boldsymbol{X}_i = \boldsymbol{x}]}_{\text{CATE estimation error}}\end{aligned}$$

- Inference of heterogenous treatment effects depends on
  1. How predictive $\boldsymbol{X}_i$ is of τ_i
  2. How good your model is for estimating $\tau(\boldsymbol{x})$

Various CATE Estimation Strategies

- *S*-learner
  1. estimate $\mu_t(\boldsymbol{x}) := \mathbb{E}[Y_i \mid T_i = t, \boldsymbol{X}_i = \boldsymbol{x}]$ using a single model
  2. compute $\hat{\tau}(\boldsymbol{x}) = \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x})$

  $\rightsquigarrow$ modeling interactions between T_i and $\boldsymbol{X}_i$ can be challenging
- *T*-learner
  1. estimate $\mu_t(\boldsymbol{x})$ separately for each $t = 0, 1$
  2. compute $\hat{\tau}(\boldsymbol{x}) = \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x})$

  $\rightsquigarrow$ difficult if the treatment assignment is lopsided, $\hat{\tau}(\boldsymbol{x})$ may not be smooth
- *X*-learner
  1. estimate $\mu_t(\boldsymbol{x})$ separately for each $t = 0, 1$
  2. impute missing potential outcomes as $\hat{\mu}_{1-T_i}(\boldsymbol{X}_i)$ and compute $\hat{\tau}_i$
  3. model estimated individual treatment effects $\hat{\tau}_i$ using $\boldsymbol{X}_i$

  $\rightsquigarrow$ more robust, but $\hat{\tau}(\boldsymbol{x})$ may not be smooth

Can we model the CATE directly?

R Learner

- Model:

$$
\begin{aligned}
Y_i(t) &= \mathbb{E}[Y_i(t) \mid \boldsymbol{X}_i] + \epsilon_i(t) \\
&= \mathbb{E}[Y_i(0) \mid \boldsymbol{X}_i] \ + \ t \times \mathbb{E}[Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i] + \ \epsilon_i(t) \quad \text{for } t = 0, 1 \\
\Longrightarrow \quad & Y_i = \mu_0(\boldsymbol{X}_i) + T_i \tau(\boldsymbol{X}_i) + \epsilon_i, \quad \text{and} \quad \mathbb{E}[Y_i \mid \boldsymbol{X}_i] = \mu_0(\boldsymbol{X}_i) + \pi(\boldsymbol{X}_i)\tau(\boldsymbol{X}_i)
\end{aligned}
$$

- Partial linear regression for (residualized) observed data:

$$
Y_i - \mathbb{E}[Y_i \mid \boldsymbol{X}_i] \ = \ \{T_i - \pi(\boldsymbol{X}_i)\}\tau(\boldsymbol{X}_i) + \epsilon_i
$$

- Estimation procedure based on cross-fitting (optional)
  1. Train models for $\pi(\boldsymbol{x})$ and $\mu(\boldsymbol{x}) = \mathbb{E}[Y_i \mid \boldsymbol{X}_i]$ and obtain $\hat{\pi}(\boldsymbol{x})$ and $\hat{\mu}(\boldsymbol{x})$ in the training sample
  2. Obtain the CATE estimate in the test sample via

$$
\hat{\tau} \ = \ \underset{\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} [\{Y_i - \hat{\mu}(\boldsymbol{X}_i)\} - \{T_i - \hat{\pi}(\boldsymbol{X}_i)\}\tau(\boldsymbol{X}_i)]^2 + \underbrace{\Lambda_n(\tau)}_{\text{regularization}}
$$

  3. Repeat the two steps by swapping the training and test samples and compute the average
- A related method: *DR*-learner

Statistical Evaluation of Heterogeneous Treatment Effects

- The above meta-learners can be used in conjunction with machine learning models
- Popular *S*-learners:
  - regression trees (e.g., CausalTree)
  - random forest (e.g., CausalForest)
  - Bayesian Additive Regression Tree (e.g., Bayesian Causal Forest)

- How can we make statistical inference for heterogeneous treatment effects discovered by a generic ML algorithm?

- Sorted Group Average Treatment Effect (GATES)

$$\tau_k \ := \ \mathbb{E}[Y_i(1) - Y_i(0) \mid c_{k-1} \leq s(\boldsymbol{X}_i) < c_k] \quad \text{for} \quad k = 1, 2, \ldots, K$$

where $s(\boldsymbol{X}_i)$ is a score (e.g., $s(\boldsymbol{X}_i) = \hat{\tau}(\boldsymbol{X}_i)$) and $c_k := \inf\{c \in \mathbb{R} : \Pr(s(\boldsymbol{X}_i) \leq c) \geq k/K\}$ is a cutoff ($p_0 = -\infty$, $p_K = \infty$)

GATES Estimation

- The difference-in-means GATES estimator:

$$\hat{\tau}_k \;=\; \frac{1}{n_{1k}}\sum_{i=1}^{n} Y_i T_i \hat{f}_k(\boldsymbol{X}_i) - \frac{1}{n_{0k}}\sum_{i=1}^{n} Y_i(1-T_i)\hat{f}_k(\boldsymbol{X}_i),$$

  where $\hat{f}_k(\boldsymbol{X}_i) \;=\; 1\{\hat{c}_{k-1} \leq s(\boldsymbol{X}_i) < \hat{c}_k\}$

- Uncertainty quantification for $\hat{\tau}_k$
  - Neyman's repeated sampling framework
  - Sources of randomness: random sampling of units, random assignment of treatment
  - This leads to the confidence interval for *fixed* $s(\boldsymbol{x})$
  - Use training sample to estimate $s(\boldsymbol{x})$ and test sample to compute $\hat{\tau}_k$ and its CI

- How should we account for the estimation uncertainty of $s(\boldsymbol{x})$?
  - cross-fitting: swap the roles of training and test samples
  - additional source of uncertainty: random splitting

Simulation Evidence (Cross-Fitting)

	$n = 100$			$n = 500$			$n = 2500$		
Estimator	bias	s.d.	coverage	bias	s.d.	coverage	bias	s.d.	coverage
Causal Forest									
$\hat{\tau}_1$	−0.053	2.971	94.0%	−0.007	1.572	95.6%	−0.007	0.594	97.7%
$\hat{\tau}_2$	−0.061	2.584	95.9	−0.038	1.075	98.2	0.011	0.541	98.6
$\hat{\tau}_3$	−0.012	2.560	96.7	−0.054	1.058	97.7	0.019	0.465	98.1
$\hat{\tau}_4$	−0.119	2.865	97.4	0.066	1.149	97.9	−0.009	0.509	98.6
$\hat{\tau}_5$	0.140	3.447	94.1	0.001	1.620	96.0	−0.006	0.620	98.3
LASSO									
$\hat{\tau}_1$	−0.125	3.196	97.6%	−0.025	1.488	96.0%	−0.004	0.669	96.0%
$\hat{\tau}_2$	0.036	2.281	97.5	−0.069	1.027	97.9	−0.019	0.590	98.9
$\hat{\tau}_3$	−0.126	2.354	96.6	−0.019	1.000	97.9	0.037	0.488	97.5
$\hat{\tau}_4$	−0.003	2.536	96.8	0.035	1.174	96.8	0.033	0.642	97.2
$\hat{\tau}_5$	0.111	3.615	96.2	0.047	1.811	95.0	0.022	0.697	95.3

Empirical Application

- National Supported Work Demonstration Program (LaLonde 1986)
- Temporary employment program to help disadvantaged workers by giving them a guaranteed job for 9 to 18 months

- Data
  - sample size: $n_1 = 297$ and $n_0 = 425$
  - outcome: annualized earnings in 1978 (36 months after the program)
  - 7 pre-treatment covariates: demographics and prior earnings

- Setup
  - S-learner: Causal Forest, BART, and Lasso
  - Sample-splitting: 2/3 of the data as training data
  - Cross-fitting: 3 folds

GATES Estimates (in 1,000 US Dollars)

	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$	$\hat{\tau}_5$
Sample-splitting					
BART	2.90	-0.73	-0.02	3.25	2.57
	$[-2.25, 8.06]$	$[-5.05, 3.58]$	$[-3.47, 3.43]$	$[-1.53, 8.03]$	$[-3.82, 8.97]$
Causal Forest	3.40	0.13	-0.85	-1.91	7.21
	$[-1.29, 3.40]$	$[-5.37, 5.63]$	$[-5.22, 3.52]$	$[-5.16, 1.34]$	$[1.22, 13.19]$
Lasso	1.86	2.62	-2.07	1.39	4.17
	$[-3.59, 7.30]$	$[-1.69, 6.93]$	$[-5.39, 1.26]$	$[-2.95, 5.73]$	$[-2.30, 10.65]$
Cross-fitting					
BART	0.40	-0.15	-0.40	2.52	2.19
	$[-3.79, 4.59]$	$[-2.54, 2.23]$	$[-3.37, 2.56]$	$[-0.99, 6.03]$	$[-0.73, 5.11]$
Causal Forest	-3.72	1.05	5.32	-2.64	4.55
	$[-6.52, -0.93]$	$[-2.28, 4.37]$	$[2.63, 8.01]$	$[-5.07, -0.22]$	$[1.14, 7.96]$
Lasso	0.65	0.45	-2.88	1.32	5.02
	$[-3.65, 4.94]$	$[-3.28, 4.18]$	$[-5.38, -0.38]$	$[-1.83, 4.48]$	$[-0.14, 10.18]$

Nonparametric Test of No Heterogeneity

- Can your ML algorithm detect heterogeneity?
- Null hypothesis of no heterogeneity:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_K$$

- Test statistic:

$$\hat{\boldsymbol{\xi}} = (\hat{\tau}_1 - \hat{\tau}, \hat{\tau}_2 - \hat{\tau}, \ldots, \hat{\tau}_K - \hat{\tau})^\top$$

  where $\hat{\tau}$ is the difference-in-means estimate of the ATE

- Asymptotic reference distribution:

$$\hat{\boldsymbol{\xi}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\xi}} \ \sim \ \chi^2_K$$

  where $\widehat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix

- We can also incorporate the estimation uncertainty of ML algorithm

Nonparametric Test of Rank-Consistency

- Can your ML algorithm correctly rank groups?
- Null hypothesis of rank consistency

$$H_0 : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$$

- Test statistic:

$$\hat{\boldsymbol{\zeta}} = \underset{\boldsymbol{\zeta}}{\operatorname{argmin}} \|\boldsymbol{\zeta} - \hat{\boldsymbol{\tau}}\|^2 \quad \text{subject to } \zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_K$$

- Asymptotic reference distribution:

$$(\hat{\boldsymbol{\zeta}} - \hat{\boldsymbol{\tau}})^\top \widehat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\zeta}} - \hat{\boldsymbol{\tau}}) \sim \bar{\chi}^2_K$$

- Again, we can incorporate the estimation uncertainty of ML algorithm

Back to the Empirical Application

	Causal Forest		BART		Lasso	
	stat	*p*-value	stat	*p*-value	stat	*p*-value
Sample-splitting						
Homogeneous Treatment Effects	9.78	0.082	2.76	0.737	5.26	0.362
Rank-consistent Treatment Effects	3.07	0.323	1.13	0.657	3.14	0.302
Cross-fitting						
Homogeneous Treatment Effects	30.29	0.000	2.32	0.803	10.79	0.056
Rank-consistent Treatment Effects	0.06	0.691	0.04	0.885	0.45	0.711

Identification of Exceptional Responders

- In the GATES estimation, the cutoff c is given
- Goal: provide a statistical guarantee when selecting c using the data
- The problem is trivial if we had an infinite amount of data

$$p^* = \underset{p\in[0,1]}{\operatorname{argmax}}\, \Psi(p) \quad \text{where } \Psi(p) \;=\; \mathbb{E}[Y_i(1) - Y_i(0) \mid \underbrace{F(S_i)}_{\text{CDF of } S_i} \geq p],$$

where $S_i = s(\boldsymbol{X}_i)$

1. sample size may not be large
2. ML estimates of CATE may be biased and noisy
3. proportion of exceptional responders may be small

- Standard method suffers from multiple testing problem:

$$\hat{p}_n \;=\; \underset{p\in[0,1]}{\operatorname{argmax}}\, \widehat{\Psi}_n(p) \quad \text{where } \widehat{\Psi}_n(p) \;=\; \underbrace{\sum_{i=1}^{\lfloor np \rfloor} \frac{T_{[n,i]}Y_{[n,i]}}{pn_1} - \frac{(1-T_{[n,i]})Y_{[n,i]}}{pn_0}}_{\text{difference-in-means estimator for top } p \text{ proportion}}$$

where $S_{[n,1]} \geq S_{[n,2]}, \ldots, \geq S_{[n,n]}$

Providing a Statistical Guarantee

- (one-sided) Uniform confidence band:

$$\mathbb{P}\left(\forall p \in [0,1],\ \Psi(p) \geq \widehat{\Psi}_n(p) - C_n(p,\alpha)\right) \geq 1-\alpha.$$

- Safe identification of exceptional responders by maximizing the lower confidence band

$$\underline{\hat{p}}_n \;=\; \operatorname*{argmax}_{p\in[0,1]} \widehat{\Psi}_n(p) - C_n(p,\alpha),$$

  implying for the true optimal p^*

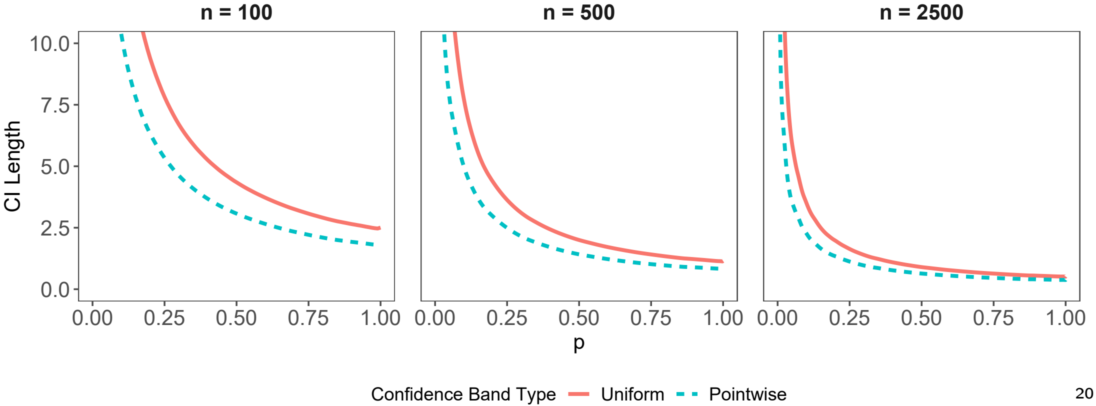
$$\begin{aligned}\mathbb{P}\left(\Psi(p^*) \geq \widehat{\Psi}_n(\underline{\hat{p}}_n) - C_n(\underline{\hat{p}}_n,\alpha)\right) &\geq \; \mathbb{P}\left(\Psi(\underline{\hat{p}}_n) \geq \widehat{\Psi}_n(\underline{\hat{p}}_n) - C_n(\underline{\hat{p}}_n,\alpha)\right) \\ &\geq \; 1-\alpha.\end{aligned}$$

- Other data-driven selection of p is possible: e.g., for a given c

$$\begin{aligned}\text{estimate}\quad \underline{\hat{p}}_n(c) \;&=\; \sup\{p\in[0,1] : \widehat{\Psi}_n(p) - C_n(p,\alpha) \geq c\}, \\ \text{to target}\quad p^*(c) \;&=\; \sup\{p\in[0,1] : \Psi(p) \geq c\}\end{aligned}$$

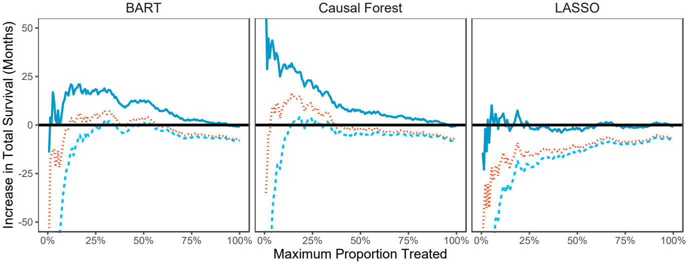
Simulation Studies

ML algorithm	Uniform			Pointwise		
	$n = 100$	$n = 500$	$n = 2500$	$n = 100$	$n = 500$	$n = 2500$
BART	96.1%	96.0%	95.2%	87.2%	76.5%	70.3%
Causal Forest	96.0%	95.3%	95.7%	83.7%	77.1%	71.9%
LASSO	95.8%	95.6%	95.6%	84.1%	76.0%	69.8%



Empirical Application

- Clinical trial data on late-stage prostate cancer ($n_1 = 125$, $n_0 = 127$)
- Outcome: total survival in months, Treatment: estrogen
- Sample-split (40% train., 60% eval.), ATE estimate -0.3 month



ML algorithm	Estimated proportion of exceptional responders	Estimated GATES	90% uniform confidence band
Causal Forest	18.8%	27.2	$(4.45, \infty)$
BART	32.2%	18.1	$(2.12, \infty)$
LASSO	91.2%	1.35	$(-6.26, \infty)$

Example R code for Sample Splitting

```r
library(evalHTE)

# sample splitting
fit_split <- estimate_hte(
  treatment = "z", form = formula, data = data, algorithms = "causal_forest",
  n_folds = 2, split_ratio = 0.5, ngates = 5, meta_learner = "slearner"
)
# evaluate HTE
est_split <- evaluate_hte(fit_split)
# summary and plot of estimated GATES
summary(est_split)
plot(est_split)

# hypothesis tests
tests_split <- test_itr(est_split, nsim = 200)
summary(test_split)
```

Example R code for Cross-Fitting

```r
# cross-fitting
fit_cv <- estimate_hte(
  treatment = "z", form = formula, data = data, algorithms = "causal_forest",
  n_folds = 5, ngates = 5, meta_learner = "slearner"
)
# evaluate HTE
est_cv <- evaluate_hte(fit_cv)
# summary and plot of estimated GATES
summary(est_cv)
plot(est_cv)

# hypothesis tests
tests_cv <- test_itr(est_cv, nsim = 200)
summary(tests_cv)
```

References

- Papers:
  1. Imai, Kosuke and Michael Lingzhi Li. (2025). "Statistical Inference for Heterogeneous Treatment Effects Discovered by Generic Machine Learning in Randomized Experiments." *Journal of Business & Economic Statistics*, Vol. 43, No. 1, pp. 256–268.
  2. Li, Michael Lingzhi and Kosuke Imai. "Statistical Performance Guarantee for Subgroup Identification with Generic Machine Learning." arXiv preprint. arXiv:2310.07973

- Open-source software:
  Li, Michael Lingzhi, and Kosuke Imai. "evalHTE: Evaluating Heterogeneous Treatment Effects." available at https://cran.r-project.org/package=evalHTE

Part II: Individualized Treatment Rules

Why Individualized Treatment Rules?

- Some benefit from the treatment while others are harmed by it
- Use individual characteristics to decide who should receive the treatment
    - personalized medicine
    - micro targeting

- Formalization of problem:

$$\hat{\pi} \; = \; \underset{\pi \in \Pi}{\operatorname{argmax}} \, \mathbb{E}[Y_i(\pi(\boldsymbol{X}_i))]$$

    - (individualized) treatment rule: $\pi : \mathcal{X} \to \{0, 1\}$
    - Policy value: $\mathbb{E}[Y_i(\pi(\boldsymbol{X}_i)]$
    - Class of individualized treatment rules: Π
        - linear policy $\pi(\boldsymbol{X}_i) = 1\{\boldsymbol{X}_i^\top \boldsymbol{\beta} > 0\}$
        - avoidance of overfitting
    - budget, fairness, and other constraints can be added

Deriving Optimal Individualized Treatment Rules

- Key identity for the policy value:

$$\begin{aligned}\mathbb{E}[Y_i(\pi(\boldsymbol{X}_i))] &= \mathbb{E}[\pi(\boldsymbol{X}_i)Y_i(1) + (1-\pi(\boldsymbol{X}_i))Y_i(0)] \\ &= \mathbb{E}[Y_i(0)] + \mathbb{E}[\pi(\boldsymbol{X}_i)\underbrace{(Y_i(1) - Y_i(0))}_{=\tau_i}] \;=\; \mathbb{E}[Y_i(0)] + \mathbb{E}[\pi(\boldsymbol{X}_i)\tau(\boldsymbol{X}_i)]\end{aligned}$$

- Oracle: assign the treatment if CATE is positive, i.e., $\pi(\boldsymbol{X}_i) = 1\{\tau(\boldsymbol{X}_i) > 0\}$
- Empirical Risk Maximization:

$$\hat{\pi} \;=\; \underset{\pi\in\Pi}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \pi(\boldsymbol{X}_i)\hat{\tau}_i$$

1. Experimental study (IPW score): $\hat{\tau}_i = \frac{Y_i T_i}{e(\boldsymbol{X}_i)} - \frac{Y_i(1-T_i)}{1-e(\boldsymbol{X}_i)}$
2. Observational study (doubly-robust score):
   $\hat{\tau}_i = \hat{\mu}_1(\boldsymbol{X}_i) - \hat{\mu}_0(\boldsymbol{X}_i) + \frac{T_i - \hat{e}(\boldsymbol{X}_i)}{\hat{e}(\boldsymbol{X}_i)(1-\hat{e}(\boldsymbol{X}_i))}(Y_i - \hat{\mu}_{T_i}(\boldsymbol{X}_i))$

Outcome Weighted Learning

- Another equality

$$\begin{aligned}\mathbb{E}[Y_i(\pi(\boldsymbol{X}_i))] &= \mathbb{E}[1\{\pi(\boldsymbol{X}_i)=1\}Y_i(1)+1\{\pi(\boldsymbol{X}_i)=0\}Y_i(0)] \\ &= \mathbb{E}[\mathbb{E}[1\{\pi(\boldsymbol{X}_i)=T_i\}Y_i \mid T_i=1,\boldsymbol{X}_i]]+\mathbb{E}[\mathbb{E}[1\{\pi(\boldsymbol{X}_i)=T_i\}Y_i \mid T_i=0,\boldsymbol{X}_i]] \\ &= \mathbb{E}\left[\frac{1\{\pi(\boldsymbol{X}_i)=T_i\}Y_i}{T_i e(\boldsymbol{X}_i)+(1-T_i)(1-e(\boldsymbol{X}_i))}\right]\end{aligned}$$

  where the second equality follows from unconfoundedness
- Equivalent minimization problem:

$$\underset{\pi\in\Pi}{\operatorname{argmax}}\,\mathbb{E}[Y_i(\pi(\boldsymbol{X}_i))] = \underset{\pi\in\Pi}{\operatorname{argmin}}\,\mathbb{E}[Y_i(1-\pi(\boldsymbol{X}_i))]$$

- Weighted classification problem

$$\underset{\pi\in\Pi}{\operatorname{argmin}}\,\frac{1}{n}\sum_{i=1}^{n}\underbrace{\frac{Y_i}{T_i e(\boldsymbol{X}_i)+(1-T_i)(1-e(\boldsymbol{X}_i))}}_{weights}\underbrace{1\{T_i\neq\pi(\boldsymbol{X}_i)\}}_{\text{misclassification}}$$

Experimental Evaluation of Fixed Individualized Treatment Rules

- Consider a fixed (for now) individualized treatment rule
  - ITR is obtained from an external dataset (e.g., sample splitting)
  - no assumption about ITR (e.g., any causal ML, heuristic rule)

- Evaluation metric examples:
  1. Policy value (population average value or PAV):
  $$\lambda_\pi \;=\; \mathbb{E}[Y_i(\pi(\boldsymbol{X}_i))]$$
  2. Population average prescriptive effect (PAPE)
  $$\gamma_\pi \;=\; \mathbb{E}[Y_i(\pi(\boldsymbol{X}_i)) - pY_i(1) - (1-p)Y_i(0)]$$
  where $p = \Pr(\pi(\boldsymbol{X}_i) = 1)$ is the proportion treated under the ITR π
  3. Difference in PAV between two ITRs
  $$\lambda_{\pi_1} - \lambda_{\pi_2}$$

Statistical Inference for the Population Average Value

- A natural estimator:

$$\hat{\lambda}_\pi = \frac{1}{n_1} \underbrace{\sum_{i=1}^{n} Y_i \pi(\boldsymbol{X}_i) T_i}_{\text{treated units who should be treated}} + \frac{1}{n_0} \underbrace{\sum_{i=1}^{n} Y_i (1-\pi(\boldsymbol{X}_i))(1-T_i)}_{\text{untreated units who should not be treated}},$$

- Sources of uncertainty: random sampling, randomized treatment assignment
- Unbiasedness: $\mathbb{E}[\hat{\lambda}_\pi] = \lambda_\pi$
- Variance:

$$\mathbb{V}(\hat{\lambda}_\pi) = \frac{\mathbb{V}[\pi(\boldsymbol{X}_i) Y_i(1)]}{n_1} + \frac{\mathbb{V}[(1-\pi(\boldsymbol{X}_i)) Y_i(0)]}{n_0}$$

  where all observations are used to estimate the variance
- Similar results for the PAPE with a negligible finite-sample bias due to estimation of the proportion treated p

Using the Same Data for Learning and Evaluation

- Cross-fitting procedure:
  1. randomly split the data into K folds: $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K$
  2. learn an ITR using $K-1$ folds: $\hat{\pi}_{-k}$ (further split may be necessary)
  3. evaluate it with the held-out set: $\hat{\lambda}_{\hat{\pi}_{-k}}(\boldsymbol{Z}_k)$
  4. repeat the process for each k and compute an average
- Additional source of uncertainty: random splitting
- ML algorithm:

$$F : \mathcal{Z} \longrightarrow \Pi$$

  where $\boldsymbol{Z}_{\text{train}} \in \mathcal{Z}$ and $\hat{\pi} = F(\boldsymbol{Z}_{\text{train}}) \in \Pi$
- Estimand and unbiased estimator:

$$\lambda_F \;=\; \underbrace{\mathbb{E}[Y_i(\hat{\pi}_{Z_{\text{train}}}(\boldsymbol{X}_i))]}_{\text{average performance of } F}\;, \quad \hat{\lambda}_F \;=\; \frac{1}{K}\sum_{k=1}^{K} \hat{\lambda}_{\hat{\pi}_{-k}}(\boldsymbol{Z}_k)$$

- It is the *average* performance of F rather than a specific ITR π
- Unbiasedness: $\mathbb{E}(\hat{\lambda}_F) = \lambda_F$

Finite-sample Variance with Cross-fitting

- Correlation due to the overlap between training and test data:

$$\mathbb{V}(\hat{\lambda}_F) \quad = \quad \frac{\mathbb{V}(\hat{\lambda}_{\hat{\pi}_{-k}}(\boldsymbol{Z}_k))}{K} + \frac{K-1}{K}\underbrace{\text{Cov}(\hat{\lambda}_{\hat{\pi}_{-k}}(\boldsymbol{Z}_k), \hat{\lambda}_{\hat{\pi}_{-k'}}(\boldsymbol{Z}_{k'}))}_{\text{dependency across folds}}$$

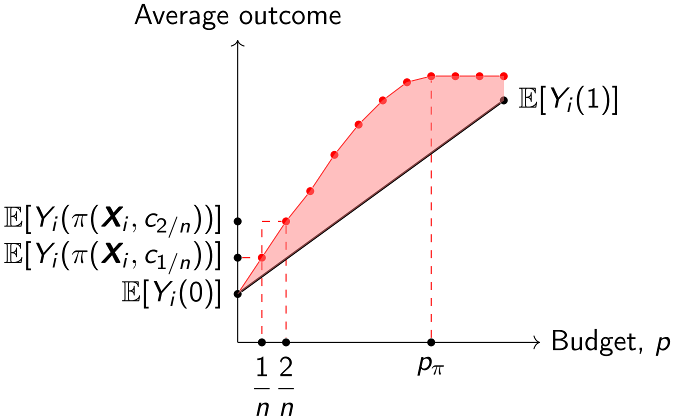
- Simplifying the expression gives:

$$\mathbb{V}(\hat{\lambda}_F) \quad = \quad \underbrace{\frac{\mathbb{V}[\hat{\pi}_{-k}(\boldsymbol{X}_i)Y_i(1)]}{n_1/K} + \frac{\mathbb{V}[(1-\hat{\pi}_{-k}(\boldsymbol{X}_i))Y_i(0)]}{n_0/K}}_{\text{variance for a fixed ITR}} - \underbrace{\frac{K-1}{K}\mathbb{E}[S_F^2]}_{\substack{\text{efficiency gain}\\ \text{due to cross-fitting}}}$$

$$+\mathbb{E}\left[\text{Cov}(\hat{\pi}_{-k}(\boldsymbol{X}_i), \hat{\pi}_{-k}(\boldsymbol{X}_j) \mid \boldsymbol{X}_i, \boldsymbol{X}_j)\tau_i\tau_j\right] \;\geq\; \mathbb{E}[S_F^2]$$

where $i \neq j$ and where S_F^2 is the sample variance of $\hat{\lambda}_{\hat{\pi}_{-k}}(\boldsymbol{Z}_k)$ across K folds

Area Under Prescriptive Effect Curve (AUPEC)



- Measure of performance across different budget constraints
- Inference is possible with or without cross-fitting
- Normalized AUPEC = average percentage gain using an ITR over the randomized treatment rule across a range of budget contraints

Simulations

- Atlantic Causal Inference Conference data analysis challenge
- Data generating process
  - 8 covariates from the Infant Health and Development Program (originally, 58 covariates and 4,302 observations)
  - population distribution = original empirical distribution
  - highly nonlinear model
- 5-fold cross fitting based on LASSO
- std. dev. for $n = 500$ is roughly half of the fixed $n = 100$ case

		$n = 100$			$n = 500$			$n = 2000$	
Estimator	cov.	bias	s.d.	cov.	bias	s.d.	cov.	bias	s.d.
Small effect									
PAV	96.9	−0.007	0.261	96.5	−0.003	0.125	97.3	0.001	0.062
PAPE	93.6	−0.000	0.171	93.0	0.000	0.093	95.3	0.001	0.041
Large effect									
PAV	96.9	−0.007	0.261	96.5	−0.003	0.125	97.3	0.001	0.062
PAPE	93.6	−0.000	0.171	93.0	0.000	0.093	95.3	0.001	0.041

Application to the STAR Experiment

- Experiment involving 7,000 students across 79 schools
- Randomized treatments (kindergarden):
  1. $T_i = 1$: small class (13–17 students)
  2. $T_i = 0$: regular class (22–25)
- Outcome: SAT scores
- 10 covariates: 4 demographic and 6 school characteristics
- Sample size: $n = 1911$, 5-fold cross-fitting

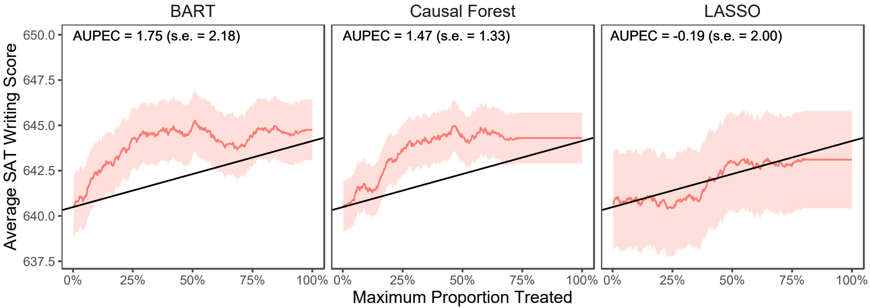
- Estimated average treatment effects:
  - SAT reading: 6.78 (s.e.=1.71)
  - SAT math: 5.78 (s.e.=1.80)
  - SAT writing:3.65 (s.e.=1.63)

Results

- ITR performance via PAPE

	BART			Causal Forest			Lasso		
	est.	s.e.	treated	est.	s.e.	treated	est.	s.e.	treated
Reading	0.19	0.37	99.3%	0.31	0.77	86.6%	0.32	0.53	87.6%
Math	0.92	0.75	84.7	2.29	0.80	79.1	1.52	1.60	75.2
Writing	1.12	0.86	88.0	1.43	0.71	67.4	0.05	1.37	74.8

- AUPEC



Example R code for Sample Splitting

```r
library(dplyr); library(evalITR)

# specifying the outcome and treatment variables
outcomes <- "g3tlangss"
treatment <- "treatment"

# specifying the formula
user_formula <- as.formula("g3tlangss ~ treatment + gender + race + birthmonth +
  birthyear + SCHLURBN + GRDRANGE + GKENRMNT + GKFRLNCH + GKBUSED + GKWHITE ")

# estimate ITR with 70-30 split
fit <- estimate_itr(treatment = treatment, form = user_formula, data = star_data,
  algorithms = c("causal_forest"), budget = 0.2, split_ratio = 0.7)

# evaluate ITR
est <- evaluate_itr(fit)
summary(est)
```

Example R code for Cross-Fitting

```r
# estimate ITR with 3 folds
fit_cv <- estimate_itr(treatment = treatment, form = user_formula, data =
    star_data,
                algorithms = c("causal_forest"), budget = 0.2, n_folds = 3)

# evaluate ITR
est_cv <- evaluate_itr(fit_cv)
summary(est_cv)

# AUPEC curve plot; works for sample splitting too
plot(est_cv)
```

Concluding Remarks

- Causal machine learning (ML) is everywhere
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)

- Safe deployment of causal ML requires uncertainty quantification
  - Statistical evaluation of HTEs and ITRs
  - No modeling assumption, Computational efficiency
  - Applicable to any complex causal ML algorithms
  - Good small sample performance

References

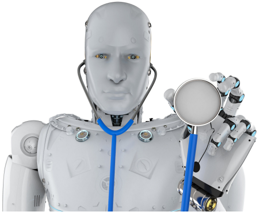
- Papers:
  Imai, Kosuke and Michael Lingzhi Li. (2023). “Experimental Evaluation of Individualized Treatment Rules.” *Journal of the American Statistical Association*, Vol. 118, No. 541, pp. 242–256.

- Open-source software:
  Li, Michael Lingzhi, and Kosuke Imai. “evalITE: Evaluating Individualized Treatment Rules.” available at https://cran.r-project.org/package=evalITR

Part III: Human-AI Collaboration

AI-assisted (Algorithm-assisted) human decision making

- AI and data-driven algorithms are everywhere in our daily lives
- But, humans still make many consequential decisions
- We have not yet outsourced high-stakes decisions to AI



  - this is true even when human decisions can be suboptimal
  - we may want to hold *someone*, rather than *something*, accountable
- Most prevalent system is AI-assisted human decision making
  - humans make decisions with the aid of AI recommendations
  - routine decisions made by individuals in daily lives
  - consequential decisions made by doctors, judges, etc.

Key questions

- How do AI recommendations influence human decisions?
    - Does AI help humans make more accurate decisions?
    - Does AI help humans improve the fairness of their decisions?

- Many have studied the accuracy and fairness of AI recommendations
    - Relatively few have researched their impacts on human decisions
    - Little is known about how AI's bias interacts with human bias

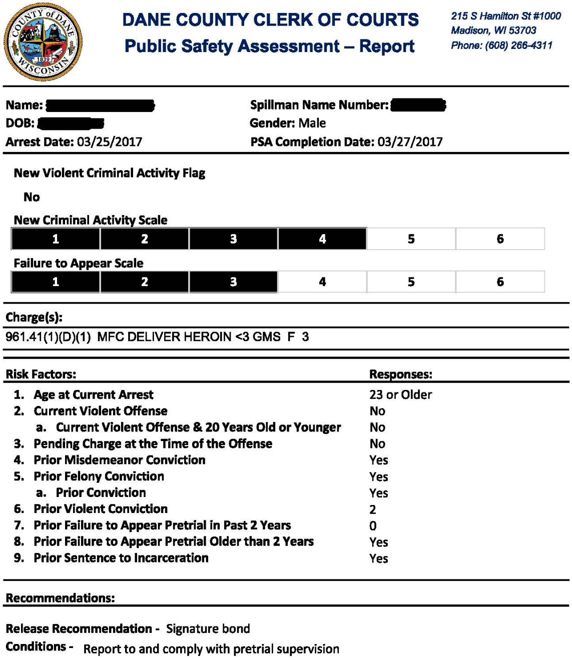
- A statistical evaluation framework for AI recommendations
    1. experimental studies: randomize human-alone vs. human+AI decisions
    2. observational studies: also applicable under unconfoundedness
    3. methodology:
        - compare human-alone, human+AI, and AI-alone
        - optimally combine human decisions with AI recommendations
    4. first ever field experiment: evaluating pretrial public safety assessment

Pretrial public safety assessment (PSA)

- AI recommendations often used in US criminal justice system
- At the first appearance hearing, judges primarily make two decisions
  1. whether to release an arrestee pending disposition of criminal charges
  2. what conditions (e.g., bail and monitoring) to impose if released
- Goal: avoid predispositional incarceration while preserving public safety
- Judges are required to consider three risk factors along with others
  1. arrestee may fail to appear in court (FTA)
  2. arrestee may engage in new criminal activity (NCA)
  3. arrestee may engage in new violent criminal activity (NVCA)
- PSA as an AI recommendation to judges
  - classifying arrestees according to FTA and NCA/NVCA risks
  - derived from an application of a machine learning algorithm to a training data set based on past observations
  - used in more than 25 states

Field experiment for evaluating the PSA

- Dane County, Wisconsin
- PSA = weighted indices of ten factors
  - age as the single demographic factor: no gender or race
  - nine factors drawn from criminal history (prior convictions and FTA)
- PSA scores and recommendation
  1. two separate ordinal six-point risk scores for FTA and NCA
  2. one binary risk score for new violent criminal activity (NVCA)
  3. aggregate recommendation: signature bond, small and large cash bail
- Judges may have other information about an arrestee
  - affidavit by a police officer about the arrest
  - defense attorney may inform about the arrestee's connections to the community (e.g., family, employment)
- Field experiment
  - PSA is calculated for each case using a computer system
  - provision of PSA is randomized across cases
  - mid-2017 – 2019 (randomization), 2-year follow-up for half sample
  - we have made the data set publicly available!



DANE COUNTY CLERK OF COURTS
Public Safety Assessment – Report

215 S Hamilton St #1000
Madison, WI 53703
Phone: (608) 266-4311

Name: [redacted]	**Spillman Name Number:** [redacted]
DOB: [redacted]	**Gender:** Male
Arrest Date: 03/25/2017	**PSA Completion Date:** 03/27/2017

New Violent Criminal Activity Flag

No

New Criminal Activity Scale

■ 1	■ 2	■ 3	■ 4	5	6

Failure to Appear Scale

■ 1	■ 2	■ 3	4	5	6

Charge(s):

961.41(1)(D)(1) MFC DELIVER HEROIN <3 GMS F 3

Risk Factors:	**Responses:**
1. Age at Current Arrest	23 or Older
2. Current Violent Offense	No
a. Current Violent Offense & 20 Years Old or Younger	No
3. Pending Charge at the Time of the Offense	No
4. Prior Misdemeanor Conviction	Yes
5. Prior Felony Conviction	Yes
a. Prior Conviction	Yes
6. Prior Violent Conviction	2
7. Prior Failure to Appear Pretrial in Past 2 Years	0
8. Prior Failure to Appear Pretrial Older than 2 Years	Yes
9. Prior Sentence to Incarceration	Yes

Recommendations:

Release Recommendation - Signature bond

Conditions - Report to and comply with pretrial supervision

Does the judge agree with PSA?

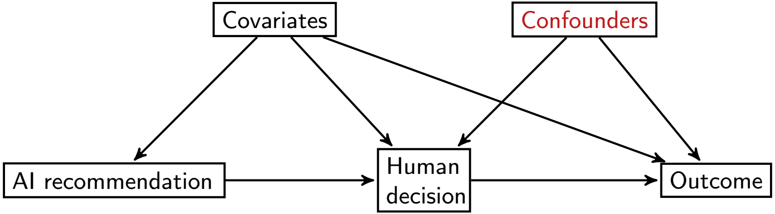
		PSA	
		Signature bond	Cash bail
Human	Signature bond	54.1% (510)	20.7 (195)
	Cash bail	9.4 (89)	15.8 (149)

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

- PSA statistically significantly influence the judge's decision, but how?

Experimental design

- Two key design features about treatment assignment:
  1. randomization (strong ignorability): human-alone vs. human+AI
  2. single blinded treatment: AI recommendations affect the outcome only through human decisions
- The proposed design is widely applicable even when stakes are high



Required assumptions

- Notation
  - AI recommendation provision (PSA or not): $Z_i \in \{0, 1\}$
  - Human decision (signature bond vs. cash bail): $D_i \in \{0, 1\}$
  - Observed outcome (FTA, NCA, or NVCA): $Y_i \in \{0, 1\}$
  - Potential decisions and outcomes: $D_i(z)$, $Y_i(z, D_i(z))$
- Assumptions
  1. Single-blinded treatment:

     $$Y_i(z, D_i(z)) = Y_i(D_i(z)) \quad \text{for all } i \text{ and } z = 0, 1$$

  2. Unconfounded treatment:

     $$Z_i \perp\!\!\!\perp \{A_i, D_i(0), D_i(1), Y_i(0), Y_i(1)\} \mid X_i \quad \text{for all } i$$

  3. Overlap: $0 < \Pr(Z_i = 1 \mid X_i = x) < 1$ for all x
- These assumptions can be guaranteed by the experimental design
- No other assumptions are required

Classification ability of decision-making system

		Decision	
		Negative ($D^* = 0$)	Positive ($D^* = 1$)
Outcome	Negative ($Y(0) = 0$)	True Negative (TN)	False Positive (FP)
	Positive ($Y(0) = 1$)	False Negative (FN)	True Positive (TP)

- (Generic) Decision D^*
  - Positive: cash bail
  - Negative: signature bond
- Outcome under release $Y(0)$
  - Positive: NCA
  - Negative: no NCA
- Classification ability measures
  - False Positive (FP): unnecessary cash bail
  - False Negative (FN): signature bond followed by NCA
- We focus on $Y(0)$ and ignore $Y(1)$

Classification risk

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

- Assign a (possibly asymmetric) 'loss' to each classification outcome
- Classification risk of decision-making system D^*

$$R(\ell_{01}; D^*) \ := \ \underbrace{\ell_{10}}_{=1} \cdot \underbrace{p_{10}(D^*)}_{\text{FNP}} + \ell_{01} \cdot \underbrace{p_{01}(D^*)}_{\text{FPP}},$$

where $p_{yd}(D^*) = \Pr(Y(0) = y, D^* = d)$ for $y, d \in \{0, 1\}$

- misclassification rate: $R(1; D^*) = \text{FNP} + \text{FPP}$

Comparing human decisions with and without AI

- Risk difference:

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

- Selective labels problem: we do not observe $Y(0)$ when $D = 1$
- FNP is identifiable but FPP is unidentified

- The difference of FPP is identifiable
  - by randomization $\Pr(Y(0) = 0 \mid Z = 1, X = x) = \Pr(Y(0) = 0 \mid Z = 0, X = x)$
  - by law of total probability

$$
\begin{aligned}
& p_{01}(D(1) \mid X = x) + p_{00}(D(1) \mid X = x) \\
= & p_{01}(D(0) \mid X = x) + p_{00}(D(0) \mid X = x)
\end{aligned}
$$

Doubly robust estimation

- Identification formula:

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

- Compound outcome: $W_i := Y_i(1-D_i) - \ell_{01}(1-Y_i)(1-D_i)$
- Three models:
  1. propensity score: $e(z,x) := \Pr(Z=z \mid X=x)$
  2. decision model: $m^D(z,x) := \Pr(D=1 \mid Z=z, X=x)$
  3. outcome model: $m^Y(z,x) := \Pr(Y=1 \mid D=0, Z=z, X=x)$

- AIPW estimator:

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

  where $\widehat{\varphi}_z(Z, X, D, Y; \ell_{01})$ is the (uncentered) influence function:

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

- Properties:
  - asymptotic normality
  - double robustness: (outcome model + decision model) × propensity score model

When do you prefer human-alone vs. human+AI?

- Hypothesis test given the relative loss ℓ_{01}:

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

- Invert this test to obtain a confidence interval on ℓ_{01}
  1. Reject H_0: prefer Human+AI over Human-alone
  2. Reject H_1: prefer Human-alone over Human+AI
  3. Fail to reject either hypothesis: statistically ambiguous

Comparing AI decisions with human-alone and human+AI

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

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

where

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

- Derive the sharp bound of risk difference: e.g., $R_{\mathrm{AI}}(\ell_{01}) - R_{\mathrm{Human}}(\ell_{01})$
- The bound width depends on the agreement between Human and AI:

$$(1 + \ell_{01})\mathbb{E}\left\{\Pr(A = 0 \mid X) - \max_{z'} \Pr(Y = 1, D = 0, A = 0 \mid Z = z', X) \right.$$
$$\left. - \max_{z'} \Pr(Y = 0, D = 0, A = 0 \mid Z = z', X)\right\}$$

- Applicable to any generic AI or any other decision system

Doubly robust estimation

- Estimation of bounds is complex
  1. data-driven choice of z'
  2. estimation of the bounds given the optimal choice of z'
- Decision and outcome models:
  1. $m^D(z, x, a) := \Pr(D = 1 \mid Z = z, X = x, A = a)$
  2. $m^Y(z, x, a) := \Pr(Y = 1 \mid D = 0, Z = z, X = x, A = a)$
- Nuisance classifier for the lower bound:

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

  - assume that this nuisance classifier is well separated
  - plug-in estimation
- Compound outcomes: $Y(1 - D)(1 - A)$, $(1 - Y)(1 - D)(1 - A)$, $(1 - A)D$, and $A(1 - D)$
- AIPW: asymptotic normality, double-robustness

When do you prefer Ai-alone vs. Human-alone?

- Same hypothesis testing framework as before:

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

- Due to partial identification, we instead test
  1. $H_{L0} : L_0 \leq 0$ vs. $H_{L1} : L_0 > 0$
  2. $H_{U0} : U_0 \geq 0$ vs. $H_{U1} : U_0 < 0$

- As before, we invert these hypothesis tests
  1. Rejecting H_{L0} implies Human is preferred over AI
  2. Rejecting H_{U0} implies AI is preferred over Human
  3. Ambiguous otherwise

Learning when to provide AI recommendations

- Policy: $\pi : \mathcal{X} \to \{0, 1\}$, provide AI recommendation or not
- Optimal policy:

$$\pi^*_{\text{rec}} \in \operatorname*{argmin}_{\pi \in \Pi} \underbrace{p_{10}(D(\pi(X))) + \ell_{01} p_{01}(D(\pi(X)))}_{= R_{\text{rec}}(\ell_{01}; \pi)}$$

  where

$$\begin{aligned} & R_{\text{rec}}(\ell_{01}; \pi) \\ = \; & R_{\text{human}}(\ell_{01}) + \mathbb{E}\left[\pi(X) \{p_{10}(D(1) \mid X) - p_{10}(D(0) \mid X)\right. \\ & \left. \qquad -\ell_{01} \cdot (p_{00}(D(1) \mid X) - p_{00}(D(0) \mid X))\}\right] \end{aligned}$$

- Empirical risk minimization using the doubly robust score

Learning when to follow AI recommendations

- Optimally following the AI recommendations (when we know the AI-alone system is better than the human decision-maker):

$$\pi^*_{\text{dec}} \in \underset{\pi \in \Pi}{\text{argmin}}\; p_{10}(\widetilde{D}(\pi(X))) + \ell_{01} p_{01}(\widetilde{D}(\pi(X))),$$

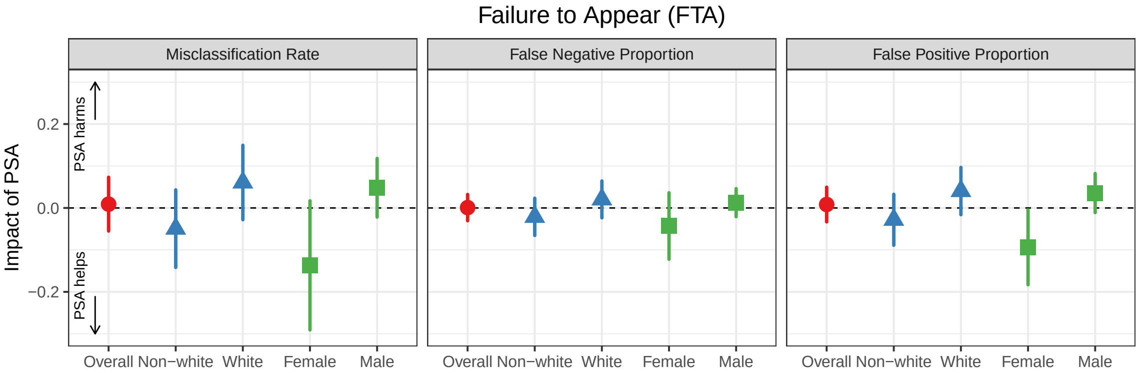
  where $\widetilde{D}(\pi(X)) = A\pi(X) + D(0)(1 - \pi(X))$

$$\begin{aligned} & R_{\text{dec}}(\ell_{01}; \pi) \\ = \; & R_{\text{human}}(\ell_{01}) + \mathbb{E}\left[\pi(X)\left\{p_{10}(A \mid X) - p_{10}(D(0) \mid X)\right.\right. \\ & \qquad\qquad\qquad\qquad \left.\left. + \ell_{01} \cdot (p_{01}(A \mid X) - p_{01}(D(0) \mid X))\right\}\right] \end{aligned}$$

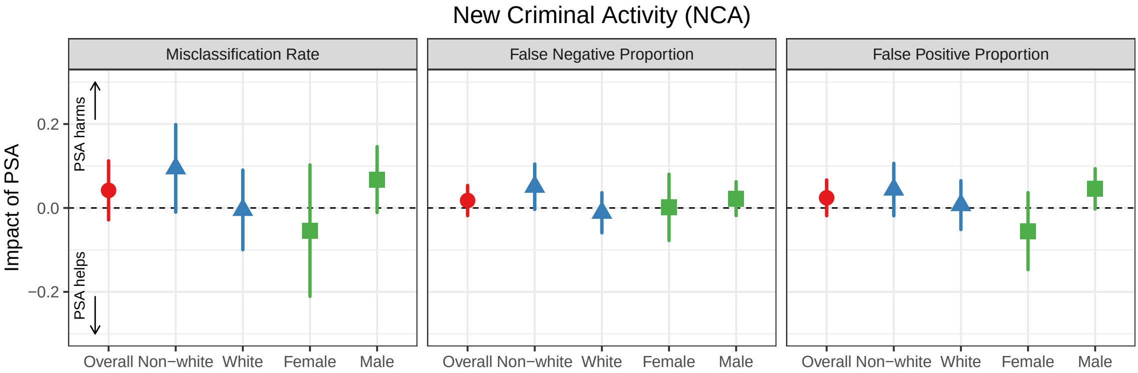
- Use the partial identification and doubly-robust score to optimize the empirical worst-case risk (upper bound) ⇝ safe policy learning

$$\pi^*_{\text{dec}} \in \underset{\pi \in \Pi}{\text{argmin}}\; \mathbb{E}[\pi(X) U_0(X)],$$

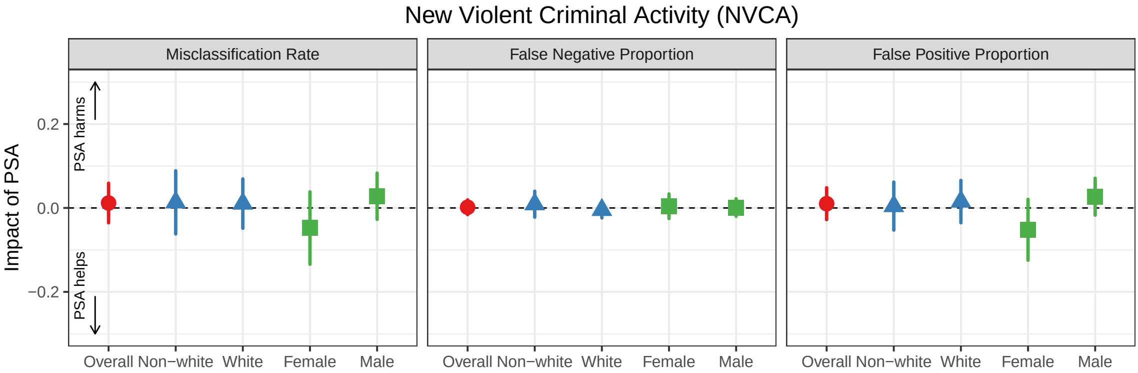
PSA recommendations do not improve human decisions



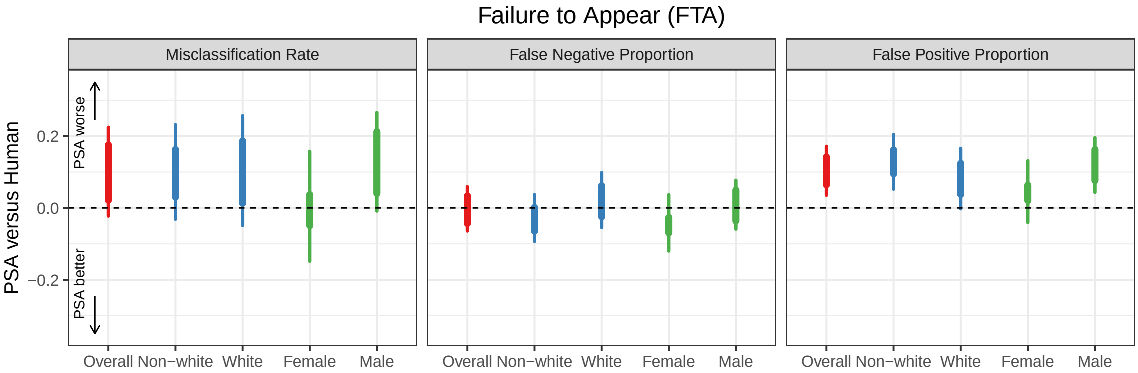
PSA recommendations do not improve human decisions



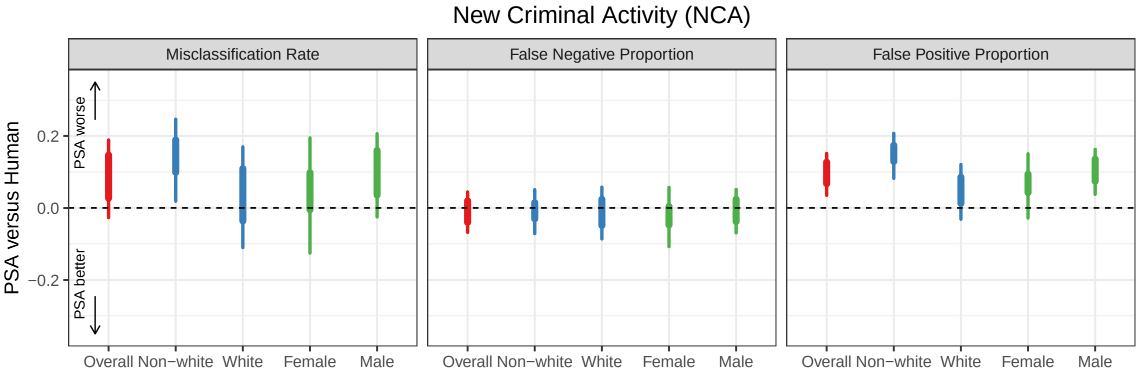
PSA recommendations do not improve human decisions



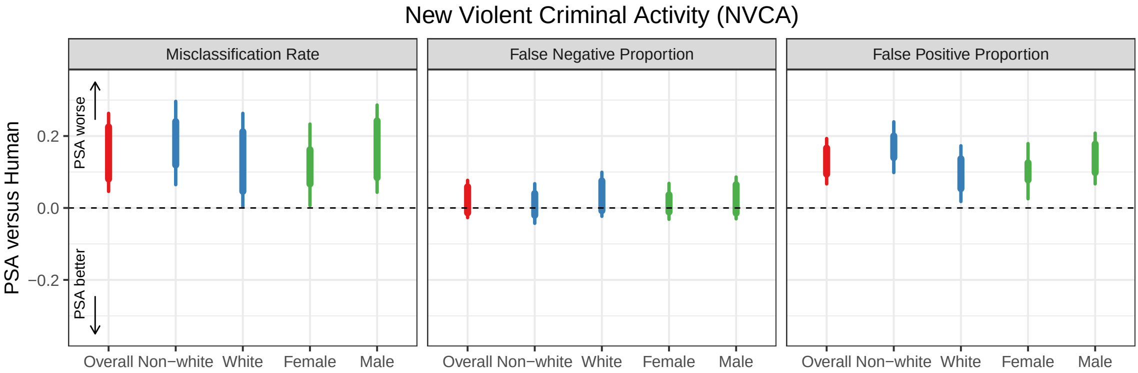
PSA-alone decisions are less accurate than human decisions



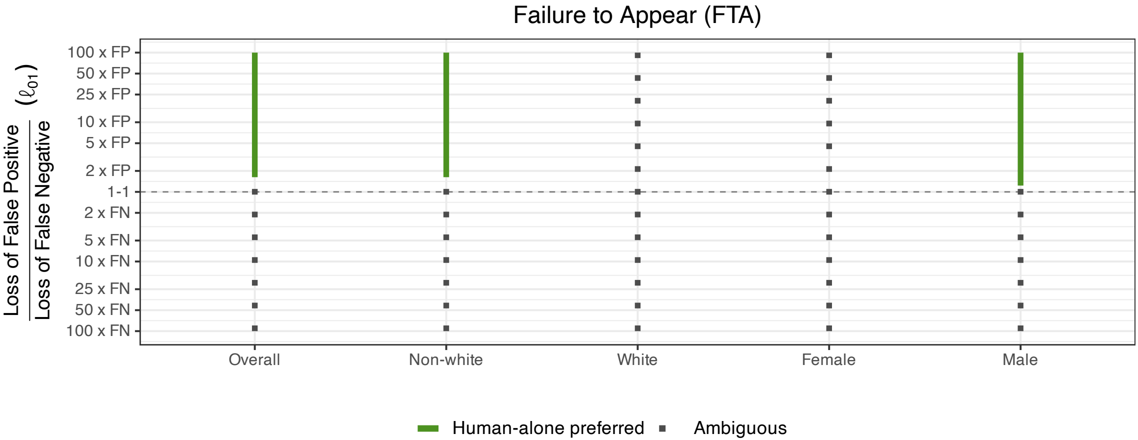
PSA-alone decisions are less accurate than human decisions



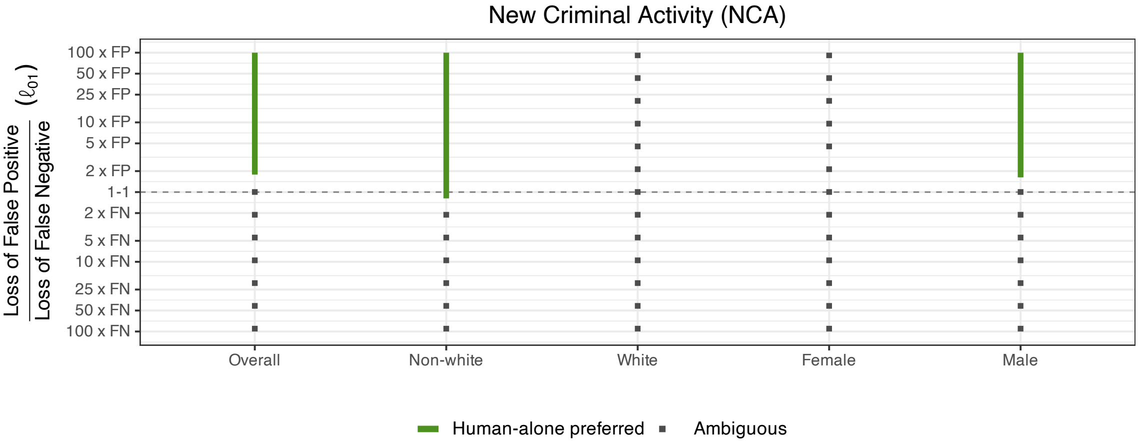
PSA-alone decisions are less accurate than human decisions



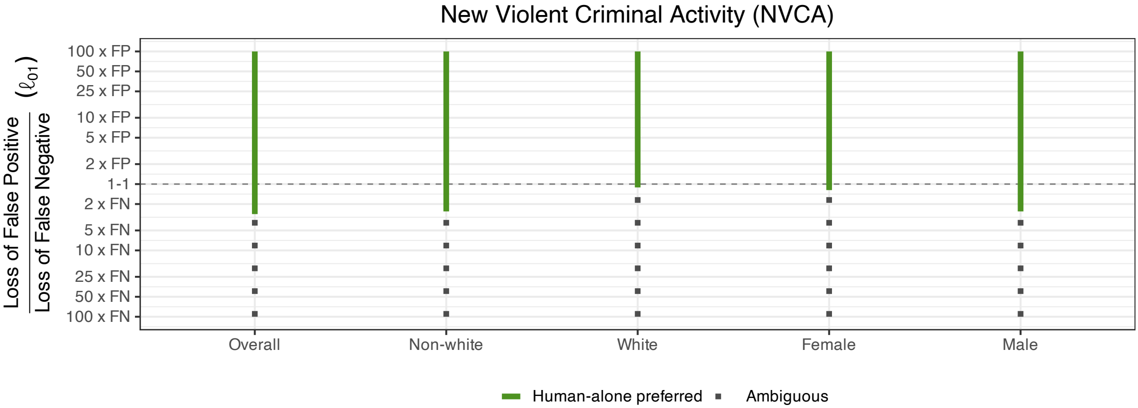
Human-alone system is preferred over PSA-alone system when the cost of false positive is high



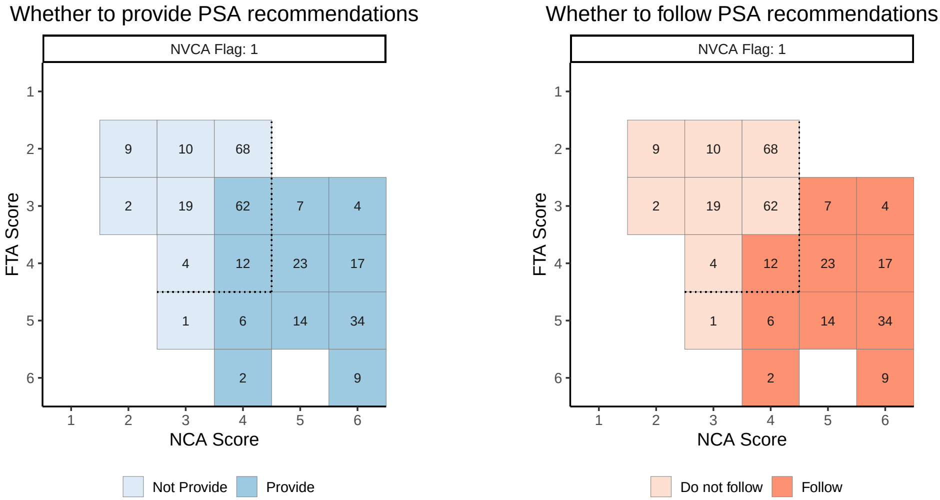
Human-alone system is preferred over AI-alone system when the cost of false positive is high



Human-alone system is preferred over AI-alone system when the cost of false positive is high



Optimally combining PSA recommendations with human decisions



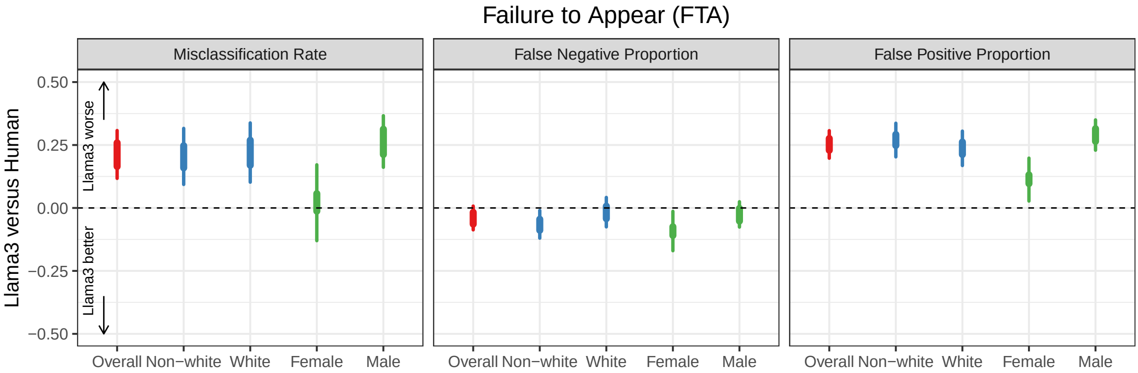
- PSA is useful only in cases with extreme recommendations

PSA is not an AI. What about the Generative AI?

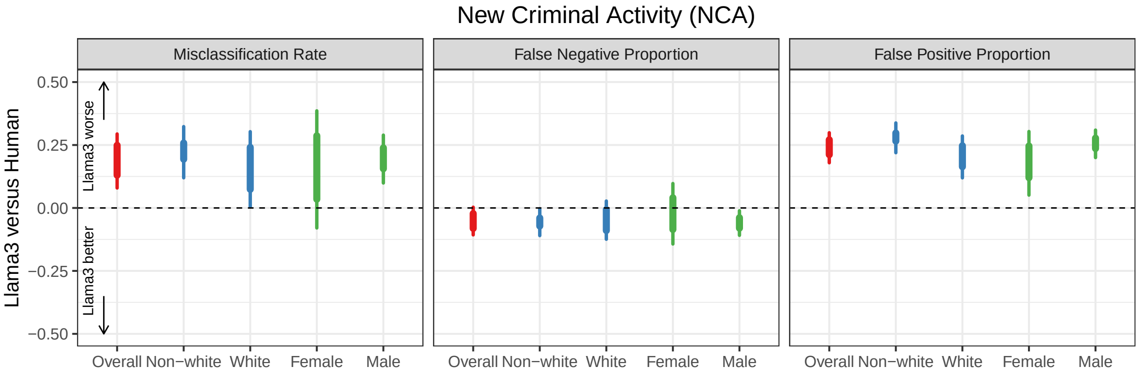
You are a judge in Dane County, Madison, Wisconsin and are asked to decide whether or not an arrestee should be released on their own recognizance or be required to post a cash bail. If you think the risk of unnecessary incarceration is too high, then the arrestee should receive own recognizance release. 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: ***[description of PSA inputs]****.*

This arrestee has the following characteristics: ***[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.*

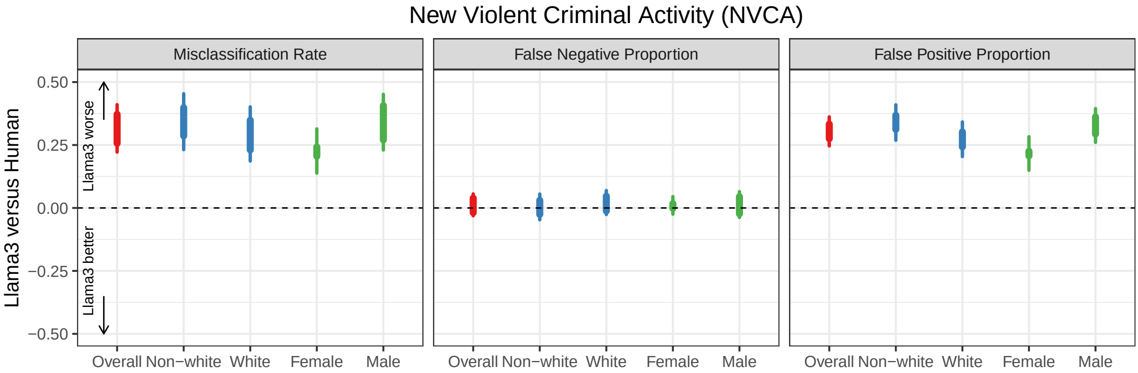
AI-alone decisions are less accurate than human decisions



AI-alone decisions are less accurate than human decisions



AI-alone decisions are less accurate than human decisions



Example R code for Sample Splitting

```r
library(aihuman)

# randomized PSA provision (0: none, 1: provided)
Z <- aihuman::NCAdata$Z
# judge's release decision (0: signature, 1: cash)
D <- if_else(aihuman::NCAdata$D == 0, 0, 1)
# dichotomized pretrial public safety assessment scores (0: signature, 1: cash)
A <- aihuman::PSAdata$DMF
# new criminal activity (0: no, 1: yes)
Y <- aihuman::NCAdata$Y

# nuisance function estimated with GBM; cov_mat being the matrix of covariates
nuis_func <- compute_nuisance_functions(Y = Y, D = D, Z = Z, V = cov_mat,
    shrinkage = 0.01, n.trees = 1000)
# estimate quantities of interest
compute_stats_aipw(Y = Y, D = D, Z = Z, nuis_funcs = nuis_func, true.pscore =
    rep(0.5, length(D)), X = NULL, l01 = 1)
```

Concluding Remarks

- We propose a methodological framework for evaluating three decision-making systems:
  1. Human-alone
  2. Human+AI
  3. AI-alone
- The proposed methodological framework is widely applicable
  - single-blinded treatment assignment is easy to implement
  - unconfoundedness + overlap enable RCT and observational studies
  - do not require AI-alone treatment condition
  - no additional assumption is required
  - open-source R software package aihuman is available
- We conducted and analyzed an RCT that evaluates the pretrial risk assessment instrument (PSA-DMF sytem):
  1. PSA recommendations do not improve human decisions
  2. Only extreme PSA recommendations are useful
  3. Both PSA and AI decisions perform worse than human decisions

References

- Paper:
  Ben-Michael, Eli, D. James Greiner, Melody Huang, Kosuke Imai, Zhichao Jiang, Sooahn Shin. (2025). "Does AI help humans make better decisions? A statistical evaluation framework for experimental and observational studies." *Proceedings of the National Academy of Sciences*, Vol. 122, No. 38, e2505106122.

- Open-source software:
  Shin, Sooahn, Zhichao Jiang, and Kosuke Imai (2024). "aihuman: Experimental Evaluation of Algorithm-Assisted Human Decision-Making." available at https://cran.r-project.org/package=aihuman