Triage Score:
A Counterfactual Risk Assessment Instrument

Kosuke Imai

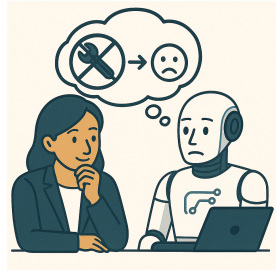
Harvard University

Joint work with Sooahn Shin, D. James Greiner, and Ryan Halen

Bridging Prediction and Intervention Problems in Social Systems
Simons Institute for the Theory of Computing
January 12, 2026

Algorithm-assisted Human Decision Making

- AI/machine learning based recommendations
  - Routine decisions made by individuals in daily lives
  - Consequential decisions made by clinicians, judges, etc.
- Risk assessment instruments ("risk scores")
  - Predicts the likelihood of an undesired outcome if *no intervention* is made
  - High risk $\rightarrow$ recommend intervention
  - Example: medicine, criminal justice system
- Methodological literature
  - Coston et al. (2020); Rambachan et al. (2022); Ben-Michael et al. (2025)
  - selective labels problem



Background on Triage Scores

- Effective decision-making requires consideration of how decisions affect outcomes
  - risk scores: only consider likelihood of an undesired outcome under no intervention
  - triage scores: incorporate counterfactual outcomes under alternative decisions

- *Triage*: *trier* or "sort" in French

  - Napoleon's military surgeon: Baron Dominique-Jean Larrey
  - prioritize worst wounds rather than first-come first-serve
  - do not fully consider counterfactual outcomes

- Modern triage systems: prioritize those who would benefit most to maximize survival given resource constraints

Triage Scores

- **Statistical decision-theoretic framework**
  - Standard decision theory: specify utility for each decision and its consequence
  - Counterfactual decision theory: incorporate counterfactual outcomes under alternative decisions when specifying each utility
  - Nuanced notions: regret, difficulty, do-no-harm, fairness etc.
    (e.g., Imai and Jiang 2023; Mueller and Pearl 2023; Ben-Michael et al. 2024; Christy and Kowalski 2024; Koch and Imai 2025)
- Two axes of comparisons
  1. Risk v.s. Triage scores
  2. Standard v.s. Counterfactual decision theory

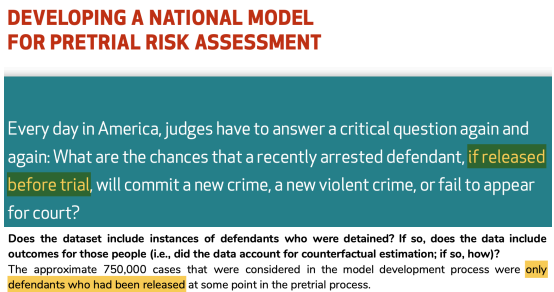
- Evaluation of decision making systems
  - Do AI recommendations help humans make better decisions?
  - When should human decision makers ignore or follow AI recommendations?
  - What is the optimal decision rule that maximizes policy maker's expected utility?

Empirical Application: Public Safety Assessment (PSA)

At the first appearance hearing, judges must decide

1. whether to release an arrestee pending disposition of charges
2. what conditions (e.g., bail and monitoring) to impose if released

- Public Safety Assessment (PSA) provides algorithmic recommendations to judges
  - Inputs: age, criminal history
  - No gender or race is used
  - Outputs: Risk factors $\rightsquigarrow$ recommendation
  - Widely used system across 25 US states
- PSA is a risk score: prediction under release
  - (1) failure to appear (FTA)
  - (2) new criminal activity (NCA)
  - (3) new *violent* criminal activity (NVCA)



Sources: Arnold Ventures report; Risk Assessment Factsheet (Stanford Law School)

Public Safety Assessment Report

(Date Created: 08/04/2018 03:36:34)

SID:	[redacted]	**Name:**	[redacted]	**Gender:**	M
PC ID:	[redacted]	**Arrest Date:**	08/04/2018	**Birth Date:**	[redacted]

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):

58-37A-5(1)(A)	USE OR POSSESSION OF DRUG PARAPHERNALIA
58-37-8(2)(H)(III)	CONTROLLED SUBSTANCE SCHEDULE III, IV, V
76-6-602	RETAIL THEFT (SHOPLIFTING)

Risk Factors:

1.	Age at Current Arrest	25
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 Offense	4
7.	Prior Failure to Appear in Past 2 Years	7
8.	Prior Failure to Appear Older Than 2 Years	Yes
9.	Prior Sentence to Incarceration	Yes

Recommendations:

1. Release with PRL4 Conditions - Maximum conditions where available: PRL3 conditions, electronic monitoring, home detention, and/or drug testing

Does a Judge Follow PSA?

- We conducted an RCT in Utah (We made the Wisconsin data publicly available)
- Experimental samples: 7K arrest cases
- Provision of PSA is randomized across cases
- PSA provision influenced the judge's decision (more agreement between judges and PSA)
  - cash bail
  - release on their own recognizance (ROR)
- Next step: evaluate and improve decision-making systems
  - Does PSA help the judge make better decisions?
  - What is the optimal decision under a given utility function?

Human

	PSA: ROR	PSA: Cash bail
Human: ROR	13.3% (609)	4.9 (224)
Human: Cash bail	53.3 (2439)	28.4 (1301)

Human+PSA

	PSA: ROR	PSA: Cash bail
Human+PSA: ROR	18.3% (807)	5.4 (240)
Human+PSA: Cash bail	46.4 (2046)	29.9 (1320)

Statistical Decision-Theoretic Framework

- Binary decision, binary outcome (generalizable to categorical cases)
- $D^* \in \{0, 1\}$: decision made by a decision-maker (e.g., human, human+PSA)
  - $D^* = 1$ if judge imposes cash bail
  - $D^* = 0$ if release on own recognizance
- $Y \in \{0, 1\}$: indicator of an undesirable outcome
  - $Y(d)$: potential outcome under decision $D^* = d$
  - $Y = 1$ if rearrested with an NCA, and $Y = 0$ otherwise
- **Principal strata**: $(Y(0), Y(1)) = (y_0, y_1)$

$$(Y(0), Y(1)) = \begin{cases} (0,0) & \text{“Safe”: no new crime under either decision} \\ (0,1) & \text{“Backlash”: new crime only under cash bail} \\ (1,0) & \text{“Preventable”: new crime only if released} \\ (1,1) & \text{“Risky”: new crime regardless of decision} \end{cases}$$

- **Risk score** considers $Y(0)$: two latent groups $\{$safe, backlash$\}$ and $\{$preventable, risky$\}$
- **Triage score** considers $(Y(0), Y(1))$: four latent groups

Evaluating Decision-Making Systems with Triage Score

		Decision	
		Release ($D^* = 0$)	Cash bail ($D^* = 1$)
Principal Strata	Safe ($Y(0) = 0,\ Y(1) = 0$)	$u^{\text{ror}}_{\neg\text{crime}} + \tilde{u}^{\text{cash}}_{\neg\text{crime}}$	$u^{\text{cash}}_{\neg\text{crime}} + \tilde{u}^{\text{ror}}_{\neg\text{crime}}$
	Backlash ($Y(0) = 0,\ Y(1) = 1$)	$u^{\text{ror}}_{\neg\text{crime}} + \tilde{u}^{\text{cash}}_{\text{crime}}$	$u^{\text{cash}}_{\text{crime}} + \tilde{u}^{\text{ror}}_{\neg\text{crime}}$
	Preventable ($Y(0) = 1,\ Y(1) = 0$)	$u^{\text{ror}}_{\text{crime}} + \tilde{u}^{\text{cash}}_{\neg\text{crime}}$	$u^{\text{cash}}_{\neg\text{crime}} + \tilde{u}^{\text{ror}}_{\text{crime}}$
	Risky ($Y(0) = 1,\ Y(1) = 1$)	$u^{\text{ror}}_{\text{crime}} + \tilde{u}^{\text{cash}}_{\text{crime}}$	$u^{\text{cash}}_{\text{crime}} + \tilde{u}^{\text{ror}}_{\text{crime}}$

- $u^{d}_{y_d}$: standard utility under decision d with outcome $Y(d) = y_d$
- $\tilde{u}^{1-d}_{y_{1-d}}$: counterfactual utility under alternative decision $1 - d$ with outcome $Y(1-d) = y_{1-d}$

Evaluating Decision-Making Systems under Risk Score

			Decision: Release ($D^* = 0$)	Decision: Cash bail ($D^* = 1$)
Baseline Outcome	$Y(0) = 0$	Safe ($Y(0) = 0,\ Y(1) = 0$)	$u_{\text{ror},\neg\text{crime}}$	$u_{\text{cash},\neg\text{crime}}$
		Backlash ($Y(0) = 0,\ Y(1) = 1$)		
	$Y(0) = 1$	Preventable ($Y(0) = 1,\ Y(1) = 0$)	$u_{\text{ror,crime}}$	$u_{\text{cash,crime}}$
		Risky ($Y(0) = 1,\ Y(1) = 1$)		

		Decision: Release ($D^* = 0$)	Decision: Cash bail ($D^* = 1$)
$Y(0) = 0$	Safe ($Y(0) = 0,\ Y(1) = 0$)	$u_{\text{ror},\neg\text{crime}} = u^{\text{ror}}_{\neg\text{crime}} + \tilde{u}^{\text{cash}}_{\neg\text{crime}}$	$u_{\text{cash},\neg\text{crime}} = u^{\text{cash}}_{\neg\text{crime}} + \tilde{u}^{\text{ror}}_{\neg\text{crime}}$
	Backlash ($Y(0) = 0,\ Y(1) = 1$)	$= u^{\text{ror}}_{\neg\text{crime}} + \tilde{u}^{\text{cash}}_{\neg\text{crime}}$	$= u^{\text{cash}}_{\neg\text{crime}} + \tilde{u}^{\text{ror}}_{\neg\text{crime}}$

Nonparametric Identification of Expected Counterfactual Utility

- Expected counterfactual utility is generally unidentifiable
  - joint potential outcomes are not simultaneously observed

$$\mathbb{E}[U(u; D^*)] = \sum_{d=0}^{K_D-1} \sum_{y_0=0}^{K_Y-1} \cdots \sum_{y_{K_D-1}=0}^{K_Y-1} u(d, \{y_d\}_{d=0}^{K_D-1}) \Pr(D^* = d, Y(0) = y_0, \ldots Y(K_D-1) = y_{K_D-1})$$

  - partial identification approaches (Li and Pearl 2019; Ben-Michael et al. 2024)

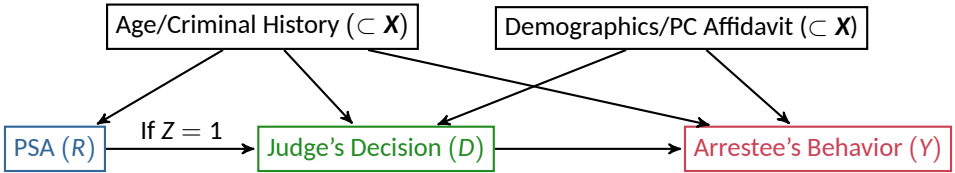
- Additive counterfactual utility (Koch and Imai, 2025):

$$u(d, \{y_d\}_{d=0}^{K_D-1}) = \underbrace{u^d_{y_d}}_{\text{standard utility}} + \sum_{d' \neq d} \underbrace{\tilde{u}^{d'}_{y_{d'}}}_{\text{counterfactual utility}} .$$

  - standard utility as a special case
  - can be a function of covariates $\boldsymbol{X}$

- Additivity as a necessary and sufficient condition for nonparametric identification under unconfoundedness

Assumptions in Our Empirical Application

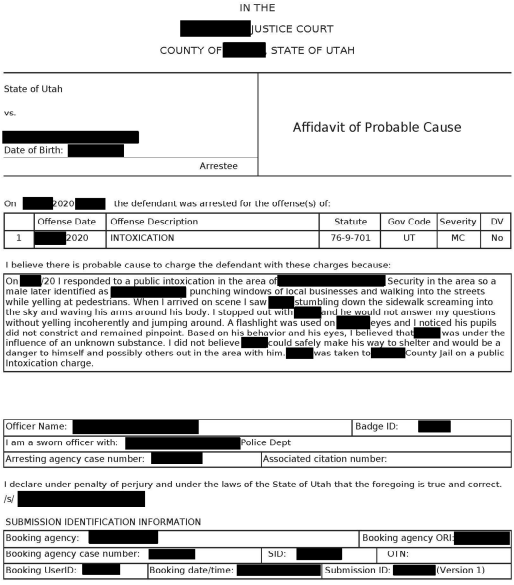


- **Assumption 1**: Single-blinded and unconfounded treatment assignment
  1. Single-blinded treatment: $Y_i(0, D_i(0)) = Y_i(1, D_i(1))$ if $D_i(0) = D_i(1)$
     $\rightsquigarrow$ PSA can only affect the outcome through the decision
  2. Unconfounded treatment assignment: $Z_i \perp\!\!\!\perp \{R_i, \{D_i(z), Y_i(d)\}_{z\in\{0,1\}, d\in\mathcal{D}}\} \mid \boldsymbol{X}_i$
  3. Overlap: $e(\boldsymbol{x}) := \Pr(Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}) \in [\lambda_Z, 1 - \lambda_Z]$ for $0 < \lambda_Z \leq \frac{1}{2}$
- **Assumption 2**: Unconfounded decision and overlap

$$\{Y_i(d)\}_{d\in\mathcal{D}} \perp\!\!\!\perp D_i \mid \boldsymbol{X}_i, Z_i, R_i, \quad \Pr(Y_i = y \mid D_i = d, Z_i = z, R_i = r, \boldsymbol{X}_i) \geq \lambda_D$$

  - holds if we have all the information that the judge takes into account (*and we do!*)
  - R can be confounded: $R \leftarrow U_D \rightarrow D$ and $R \leftarrow U_Y \rightarrow Y$ are allowed

Affidavit of Probable Cause and GenAI Powered Inference (GPI)



IN THE

[redacted] JUSTICE COURT

COUNTY OF [redacted] STATE OF UTAH

State of Utah

vs.

[redacted]

Date of Birth: [redacted]

Arrestee

Affidavit of Probable Cause

On [redacted]2020[redacted] the defendant was arrested for the offense(s) of:

	Offense Date	Offense Description	Statute	Gov Code	Severity	DV
1	[redacted]2020	INTOXICATION	76-9-701	UT	MC	No

I believe there is probable cause to charge the defendant with these charges because:

On [redacted]20 I responded to a public intoxication in the area of [redacted] Security in the area so a male later identified as [redacted] punching windows of local businesses and walking into the streets while yelling at pedestrians. When I arrived on scene I saw [redacted] stumbling down the sidewalk screaming into the sky and waving his arms around his body. I stopped out with [redacted] and he would not answer my questions without yelling incoherently and jumping around. A flashlight was used on [redacted] eyes and I noticed his pupils did not constrict and remained pinpoint. Based on his behavior and his eyes, I believed that [redacted] was under the influence of an unknown substance. I did not believe [redacted] could safely make his way to shelter and would be a danger to himself and possibly others out in the area with him. [redacted] was taken to [redacted] County Jail on a public intoxication charge.

Officer Name: [redacted]	Badge ID: [redacted]
I am a sworn officer with: [redacted] Police Dept	
Arresting agency case number: [redacted]	Associated citation number:

I declare under penalty of perjury and under the laws of the State of Utah that the foregoing is true and correct.

/s/ [redacted]

SUBMISSION IDENTIFICATION INFORMATION

Booking agency: [redacted]		Booking agency ORI: [redacted]	
Booking agency case number: [redacted]	SID: [redacted]	OTN:	
Booking UserID: [redacted]	Booking date/time: [redacted]	Submission ID: [redacted] (Version 1)	

- In Utah, PC Affidavit (along with PSA) contains all the information a judge has
- structured information: types of offense
- unstructured information: probable cause statement

- GPI (Imai and Nakamura 2025)
  - use open-source LLM to regenerate texts
  - obtain their internal representation
  - guranteed to contain all the information
  - leverage LLM's internal structure

Policy Evaluation and Learning

- Nonparametric identification of expected additive counterfactual utilities
  - We consider a general setting (Assumptions 1 and 2), which are plausible in our application
  - Under binary decision, Assumption 2 (unconfoundedness of decision) is not required for the comparison between human-alone and human+PSA
- AIPW estimators given the decision and outcome models
  - GPI to learn the "deconfounder" through the DragonNet neural network architecture
  - asymptotic normality, double-robustness, rate-robustness
- Statistical hypothesis test
  1. When do we prefer Human+PSA? Invert the hypothesis test

$$H_0 : \overline{U}(u; D(1)) \leq \overline{U}(u; D(0)) \text{ vs. } H_1 : \overline{U}(u; D(1)) > \overline{U}(u; D(0))$$

     for a given utility function u
  2. What is an optimal decision rule? Consider a covariate-dependent policy $\pi : \mathcal{X} \to \mathcal{D}$ and estimate optimal policy by solving the empirical utility maximization problem

Utility Specification

- Focus on "loss" aspects: cost and regret
  - Cost of decision: c^{cash}
  - Cost of outcome: $c^{\text{ror}}_{\text{crime}}$ and $c^{\text{cash}}_{\text{crime}}$
  - Regret from counterfactual outcome: $r^{\text{ror}}_{\neg\text{crime}}$ and $r^{\text{cash}}_{\neg\text{crime}}$

		Decision	
		Release ($D^* = 0$)	Cash bail ($D^* = 1$)
Principal Strata	Safe ($Y(0) = 0,\ Y(1) = 0$)	$-r^{\text{cash}}_{\neg\text{crime}}$	$-c^{\text{cash}} - r^{\text{ror}}_{\neg\text{crime}}$
	Backlash ($Y(0) = 0,\ Y(1) = 1$)	0	$-c^{\text{cash}} - c^{\text{cash}}_{\text{crime}} - r^{\text{ror}}_{\neg\text{crime}}$
	Preventable ($Y(0) = 1,\ Y(1) = 0$)	$-c^{\text{ror}}_{\text{crime}} - r^{\text{cash}}_{\neg\text{crime}}$	$-c^{\text{cash}}$
	Risky ($Y(0) = 1,\ Y(1) = 1$)	$-c^{\text{ror}}_{\text{crime}}$	$-c^{\text{cash}} - c^{\text{cash}}_{\text{crime}}$

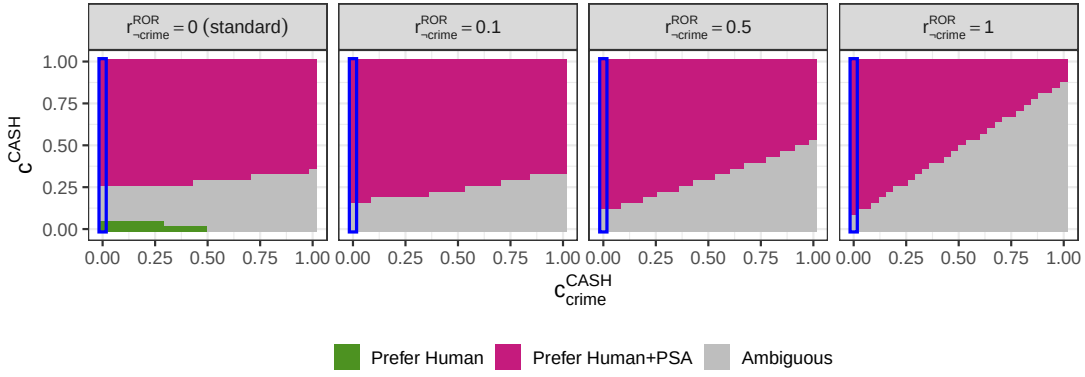
Utility Specification (cont.)

- We standardize $c^{\text{ror}}_{\text{crime}} = 1$ (cost of new crime under ROR)
  1. Standard decision theory with the risk score framework: c^{cash}
  2. Standard decision theory with the triage score framework: c^{cash}, $c^{\text{cash}}_{\text{crime}}$
  3. Counterfactual decision theory with the risk score framework: c^{cash}, $r^{\text{ror}}_{\neg\text{crime}}$
  4. Counterfactual decision theory with the triage score framework: c^{cash}, $c^{\text{cash}}_{\text{crime}}$, $r^{\text{ror}}_{\neg\text{crime}}$, $r^{\text{cash}}_{\neg\text{crime}}$
- To further reduce the number of parameters for our analysis:

$$\frac{\text{Cost of new crime under cash bail } (c^{\text{cash}}_{\text{crime}})}{\text{Cost of new crime under ROR}(c^{\text{ror}}_{\text{crime}})} = \frac{\text{Regret of no new crime under cash bail } (r^{\text{cash}}_{\neg\text{crime}})}{\text{Regret of no new crime under ROR}(r^{\text{ror}}_{\neg\text{crime}})}$$

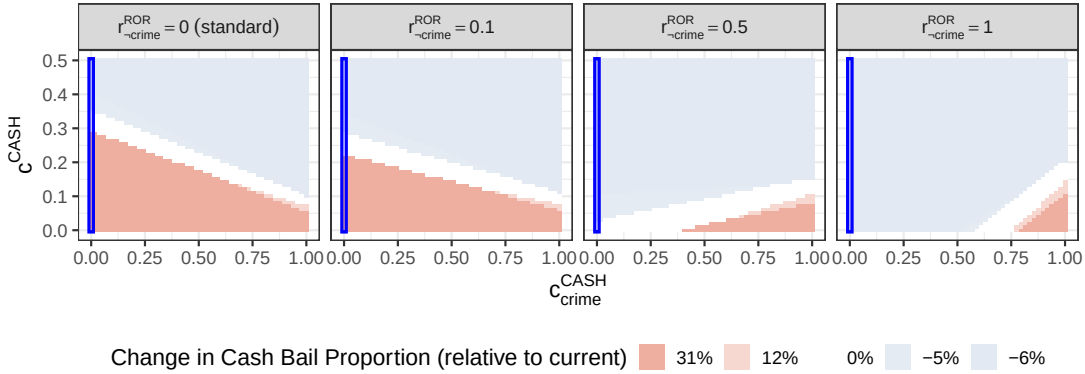
$$\rightsquigarrow r^{\text{cash}}_{\neg\text{crime}} = r^{\text{ror}}_{\neg\text{crime}} c^{\text{cash}}_{\text{crime}}$$

When do we prefer Human+PSA over Human? (Preliminary Results)



- Outcome: new criminal activity
- Blue boxes: risk score system (when $c^{\text{cash}}_{\text{crime}} = 0$)
- First panel: standard decision framework (when $r^{\text{ror}}_{\neg\text{crime}} = r^{\text{ror}}_{\text{crime}} = 0$)
- More ambiguous under triage score than risk score

Optimal Decision Rules (Preliminary Results)



- Outcome: failure to appear
- Policy class: current PSA decision rule with varying NVCA flag threshold (θ)
- PSA recommends cash bail if NVCA($\boldsymbol{X}$) $> \theta$ where NVCA($\boldsymbol{X}$) $\in [0, 7]$

- triage scores lead to more lenient optimal decisions than risk scores

Conclusion

We propose **triage score** as a comprehensive framework for decision-making

- Unlike risk scores, triage scores incorporate counterfactual outcomes and utilities
- It enables decision-makers to incorporate a wide range of ethical and practical factors
- It can be used to assess Human, Human+PSA, and PSA decision-making systems

- We apply our framework to a RCT to evaluate a pre-trial risk assessment instrument
  - Triage scores capture richer utility structures than risk scores and yield distinct results
  - Our preliminary results show that triage score framework yields higher expected utility for Human + PSA and more lenient optimal recommendations than risk score framework

- Next steps:
  - RCT with the optimal triage score (in planning!)
  - Evaluate Human+Triage system