

Triage Score: A Counterfactual Risk Assessment Instrument

Kosuke Imai

Harvard University

Mathematical Economics Seminar

Center Of Mathematical Sciences And Applications

September 22, 2026

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

Motivation: Algorithm/AI-assisted Human Decision Making

- AI/machine learning based recommendation systems

Motivation: Algorithm/AI-assisted Human Decision Making

- AI/machine learning based recommendation systems
 - Routine decisions made by individuals in daily lives

Motivation: Algorithm/AI-assisted Human Decision Making

- AI/machine learning based recommendation systems
 - Routine decisions made by individuals in daily lives
 - Consequential decisions made by clinicians, judges, etc.

Motivation: Algorithm/AI-assisted Human Decision Making

- AI/machine learning based recommendation systems
 - Routine decisions made by individuals in daily lives
 - Consequential decisions made by clinicians, judges, etc.
- Risk assessment instruments (“risk scores”)

Motivation: Algorithm/AI-assisted Human Decision Making

- AI/machine learning based recommendation systems
 - Routine decisions made by individuals in daily lives
 - Consequential decisions made by clinicians, judges, etc.
- Risk assessment instruments (“risk scores”)
 - Widely used in medicine, criminal justice system, and other areas

Motivation: Algorithm/AI-assisted Human Decision Making

- AI/machine learning based recommendation systems
 - Routine decisions made by individuals in daily lives
 - Consequential decisions made by clinicians, judges, etc.
- Risk assessment instruments (“risk scores”)
 - Widely used in medicine, criminal justice system, and other areas
 - Predicts the likelihood of an undesired outcome if *no intervention* is made

Motivation: Algorithm/AI-assisted Human Decision Making

- AI/machine learning based recommendation systems
 - Routine decisions made by individuals in daily lives
 - Consequential decisions made by clinicians, judges, etc.
- Risk assessment instruments (“risk scores”)
 - Widely used in medicine, criminal justice system, and other areas
 - Predicts the likelihood of an undesired outcome if *no intervention* is made
 - High risk → recommend intervention

Motivation: Algorithm/AI-assisted Human Decision Making

- AI/machine learning based recommendation systems
 - Routine decisions made by individuals in daily lives
 - Consequential decisions made by clinicians, judges, etc.
- Risk assessment instruments (“risk scores”)
 - Widely used in medicine, criminal justice system, and other areas
 - Predicts the likelihood of an undesired outcome if *no intervention* is made
 - High risk → recommend intervention
- We should also consider what would happen if a decision maker does *intervene*

Background on Triage Score

- Effective decision-making requires consideration of how decisions affect outcomes

Background on Triage Score

- Effective decision-making requires consideration of how decisions affect outcomes
 - risk score: only considers the likelihood of an undesired outcome under no intervention

Background on Triage Score

- Effective decision-making requires consideration of how decisions affect outcomes
 - risk score: only considers the likelihood of an undesired outcome under no intervention
 - triage score: incorporates counterfactual outcomes under alternative decisions

Background on Triage Score

- Effective decision-making requires consideration of how decisions affect outcomes
 - risk score: only considers the likelihood of an undesired outcome under no intervention
 - **triage score**: incorporates counterfactual outcomes under **alternative decisions**
- **Triage**: *trier* or “sort” in French

Background on Triage Score

- Effective decision-making requires consideration of how decisions affect outcomes
 - risk score: only considers the likelihood of an undesired outcome under no intervention
 - **triage score**: incorporates counterfactual outcomes under **alternative decisions**
- **Triage**: *trier* or “sort” in French



- Napoleon's military surgeon: Baron Dominique-Jean Larrey

Background on Triage Score

- Effective decision-making requires consideration of how decisions affect outcomes
 - risk score: only considers the likelihood of an undesired outcome under no intervention
 - **triage score**: incorporates 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

Background on Triage Score

- Effective decision-making requires consideration of how decisions affect outcomes
 - risk score: only considers the likelihood of an undesired outcome under no intervention
 - **triage score**: incorporates 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

Background on Triage Score

- Effective decision-making requires consideration of how decisions affect outcomes
 - risk score: only considers the likelihood of an undesired outcome under no intervention
 - **triage score**: incorporates 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 scores: prioritize those **who would benefit most** to maximize survival given resource constraints

Triage Score as Statistical Decision-theoretic Framework

- **Standard** decision theory: specifies utility for each decision and its consequence

Triage Score as Statistical Decision-theoretic Framework

- **Standard** decision theory: specifies utility for each decision and its consequence
- **Counterfactual** decision theory: incorporates counterfactual outcomes under alternative decisions when specifying each utility

Triage Score as Statistical Decision-theoretic Framework

- **Standard** decision theory: specifies utility for each decision and its consequence
- **Counterfactual** decision theory: incorporates counterfactual outcomes under alternative decisions when specifying each utility
- Nuanced counterfactual notions: regret, difficulty, do-no-harm, fairness etc.

Triage Score as Statistical Decision-theoretic Framework

- **Standard** decision theory: specifies utility for each decision and its consequence
- **Counterfactual** decision theory: incorporates counterfactual outcomes under alternative decisions when specifying each utility
- Nuanced counterfactual notions: regret, difficulty, do-no-harm, fairness etc.
- Two axes of comparisons

Triage Score as Statistical Decision-theoretic Framework

- **Standard** decision theory: specifies utility for each decision and its consequence
- **Counterfactual** decision theory: incorporates counterfactual outcomes under alternative decisions when specifying each utility
- Nuanced counterfactual notions: regret, difficulty, do-no-harm, fairness etc.
- Two axes of comparisons
 - ① Risk v.s. Triage scores: beyond the baseline outcome

Triage Score as Statistical Decision-theoretic Framework

- **Standard** decision theory: specifies utility for each decision and its consequence
- **Counterfactual** decision theory: incorporates counterfactual outcomes under alternative decisions when specifying each utility
- Nuanced counterfactual notions: regret, difficulty, do-no-harm, fairness etc.
- Two axes of comparisons
 - ① Risk v.s. Triage scores: beyond the baseline outcome
 - ② Standard v.s. Counterfactual decision theory: beyond the realized outcome

Triage Score as Statistical Decision-theoretic Framework

- **Standard** decision theory: specifies utility for each decision and its consequence
- **Counterfactual** decision theory: incorporates counterfactual outcomes under alternative decisions when specifying each utility
- Nuanced counterfactual notions: regret, difficulty, do-no-harm, fairness etc.
- Two axes of comparisons
 - ① Risk v.s. Triage scores: beyond the baseline outcome
 - ② Standard v.s. Counterfactual decision theory: beyond the realized outcome
- Evaluation of decision making systems in practice

Triage Score as Statistical Decision-theoretic Framework

- **Standard** decision theory: specifies utility for each decision and its consequence
- **Counterfactual** decision theory: incorporates counterfactual outcomes under alternative decisions when specifying each utility
- Nuanced counterfactual notions: regret, difficulty, do-no-harm, fairness etc.
- Two axes of comparisons
 - ① Risk v.s. Triage scores: beyond the baseline outcome
 - ② Standard v.s. Counterfactual decision theory: beyond the realized outcome
- Evaluation of decision making systems in practice
 - Do AI/algorithmic recommendations help humans make better decisions?

Triage Score as Statistical Decision-theoretic Framework

- **Standard** decision theory: specifies utility for each decision and its consequence
- **Counterfactual** decision theory: incorporates counterfactual outcomes under alternative decisions when specifying each utility
- Nuanced counterfactual notions: regret, difficulty, do-no-harm, fairness etc.
- Two axes of comparisons
 - ① Risk v.s. Triage scores: beyond the baseline outcome
 - ② Standard v.s. Counterfactual decision theory: beyond the realized outcome
- Evaluation of decision making systems in practice
 - Do AI/algorithmic recommendations help humans make better decisions?
 - 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

Empirical Application: Public Safety Assessment (PSA)

At the first appearance hearing, judges must decide

- 1 whether to release an arrestee pending disposition of charges

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

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

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

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

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

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

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

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 $\leadsto$ 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)**

DEVELOPING A NATIONAL MODEL FOR PRETRIAL RISK ASSESSMENT

Every day in America, judges have to answer a critical question again and again: What are the chances that a recently arrested defendant, **if released before trial**, will commit a new crime, a new violent crime, or fail to appear for court?

Does the dataset include instances of defendants who were detained? If so, does the data include outcomes for those people (i.e., did the data account for counterfactual estimation; if so, how)? The approximate 750,000 cases that were considered in the model development process were **only defendants who had been released** at some point in the pretrial process.

Sources: Arnold Ventures report; Risk Assessment Factsheet

Public Safety Assessment Report

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

SID:		Name:		Gender:	M
PC ID:		Arrest Date:	08/04/2018	Birth Date:	

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?

Does a Judge Follow PSA?

- We conducted an RCT in Utah (and elsewhere)

Does a Judge Follow PSA?

- We conducted an RCT in Utah (and elsewhere)
- Experimental samples: 7K arrest cases

Does a Judge Follow PSA?

- We conducted an RCT in Utah (and elsewhere)
- Experimental samples: 7K arrest cases
- Provision of PSA is randomized across cases

Does a Judge Follow PSA?

- We conducted an RCT in Utah (and elsewhere)
- 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)

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

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

Does a Judge Follow PSA?

- We conducted an RCT in Utah (and elsewhere)
- 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)
- Goal: evaluate and improve decision-making systems

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

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

Does a Judge Follow PSA?

- We conducted an RCT in Utah (and elsewhere)
- 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)
- Goal: evaluate and improve decision-making systems
 - Does PSA help the judge make better decisions?

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

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

Does a Judge Follow PSA?

- We conducted an RCT in Utah (and elsewhere)
- 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)
- Goal: evaluate and improve decision-making systems
 - Does PSA help the judge make better decisions?
 - What is the optimal decision under a given utility function?

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

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

Statistical Decision-Theoretic Framework

- Binary decision, binary outcome (generalizable to categorical cases)

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)

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 (cash)

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 (cash)
 - $D^* = 0$ if release on own recognizance (ROR)

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 (cash)
 - $D^* = 0$ if release on own recognizance (ROR)
- $Y \in \{0, 1\}$: indicator of an undesirable **outcome**

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 (cash)
 - $D^* = 0$ if release on own recognizance (ROR)
- $Y \in \{0, 1\}$: indicator of an undesirable **outcome**
 - $Y(d)$: potential outcome under decision $D^* = d$

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 (cash)
 - $D^* = 0$ if release on own recognizance (ROR)
- $Y \in \{0, 1\}$: indicator of an undesirable **outcome**
 - $Y(d)$: potential outcome under decision $D^* = d$
 - $Y = 1$ if rearrested with an NCA, and $Y = 0$ otherwise

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 (cash)
 - $D^* = 0$ if release on own recognizance (ROR)
- $Y \in \{0, 1\}$: indicator of an undesirable **outcome**
 - $Y(d)$: potential outcome under decision $D^* = d$
 - $Y = 1$ if rearrested with an NCA, and $Y = 0$ otherwise
- **Triage score** considers principal strata: $(Y(0), Y(1)) = (y_0, y_1)$

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

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 (cash)
 - $D^* = 0$ if release on own recognizance (ROR)
- $Y \in \{0, 1\}$: indicator of an undesirable **outcome**
 - $Y(d)$: potential outcome under decision $D^* = d$
 - $Y = 1$ if rearrested with an NCA, and $Y = 0$ otherwise
- **Triage score** considers principal strata: $(Y(0), Y(1)) = (y_0, y_1)$

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

- **Risk score** considers $Y(0)$: two latent groups {safe, backlash} and {preventable, hopeless}

Evaluating Decision-Making Systems with Triage Score

		Decision	
		Release ($D^* = 0$)	Cash bail ($D^* = 1$)
Principal Strata	Safe ($Y(0) = 0, Y(1) = 0$)	$u_{\neg\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\neg\text{CRIME}}^{\text{CASH}}$	$u_{\neg\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\neg\text{CRIME}}^{\text{ROR}}$
	Backlash ($Y(0) = 0, Y(1) = 1$)	$u_{\neg\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\text{CRIME}}^{\text{CASH}}$	$u_{\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\neg\text{CRIME}}^{\text{ROR}}$
	Preventable ($Y(0) = 1, Y(1) = 0$)	$u_{\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\neg\text{CRIME}}^{\text{CASH}}$	$u_{\neg\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\text{CRIME}}^{\text{ROR}}$
	Hopeless ($Y(0) = 1, Y(1) = 1$)	$u_{\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\text{CRIME}}^{\text{CASH}}$	$u_{\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\text{CRIME}}^{\text{ROR}}$

- $u_{y_d}^d$: standard utility under decision d with outcome $Y(d) = y_d$

Evaluating Decision-Making Systems with Triage Score

		Decision	
		Release ($D^* = 0$)	Cash bail ($D^* = 1$)
Principal Strata	Safe ($Y(0) = 0, Y(1) = 0$)	$u_{\neg\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\neg\text{CRIME}}^{\text{CASH}}$	$u_{\neg\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\neg\text{CRIME}}^{\text{ROR}}$
	Backlash ($Y(0) = 0, Y(1) = 1$)	$u_{\neg\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\text{CRIME}}^{\text{CASH}}$	$u_{\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\neg\text{CRIME}}^{\text{ROR}}$
	Preventable ($Y(0) = 1, Y(1) = 0$)	$u_{\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\neg\text{CRIME}}^{\text{CASH}}$	$u_{\neg\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\text{CRIME}}^{\text{ROR}}$
	Hopeless ($Y(0) = 1, Y(1) = 1$)	$u_{\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\text{CRIME}}^{\text{CASH}}$	$u_{\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\text{CRIME}}^{\text{ROR}}$

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

Evaluating Decision-Making Systems under Risk Score

		Decision	
		Release ($D^* = 0$)	Cash bail ($D^* = 1$)
Baseline Outcome	$Y(0) = 0$	$u_{\text{ROR}, \neg \text{CRIME}}$	$u_{\text{CASH}, \neg \text{CRIME}}$
	$Y(0) = 1$	$u_{\text{ROR}, \text{CRIME}}$	$u_{\text{CASH}, \text{CRIME}}$

• u_{d,y_0}

- $u_{0,y_0} = u_{y_0}^0$: standard utility under decision $d = 0$ with realized outcome $Y(0) = y_0$

Evaluating Decision-Making Systems under Risk Score

		Decision	
		Release ($D^* = 0$)	Cash bail ($D^* = 1$)
Baseline Outcome	$Y(0) = 0$	$u_{\text{ROR}, \neg \text{CRIME}}$	$u_{\text{CASH}, \neg \text{CRIME}}$
	$Y(0) = 1$	$u_{\text{ROR}, \text{CRIME}}$	$u_{\text{CASH}, \text{CRIME}}$

• u_{d,y_0}

- $u_{0,y_0} = u_{y_0}^0$: standard utility under decision $d = 0$ with realized outcome $Y(0) = y_0$
- $u_{1,y_0} = \tilde{u}_{y_0}^1$: counterfactual utility under an alternative decision $d = 1$ with counterfactual outcome $Y(0) = y_0$

Evaluating Decision-Making Systems under Risk Score

		Decision		
		Release ($D^* = 0$)	Cash bail ($D^* = 1$)	
Baseline Outcome	$Y(0) = 0$	Safe ($Y(0) = 0, Y(1) = 0$)	$u_{\text{ROR}, \neg \text{CRIME}}$ $= u_{\neg \text{CRIME}}^{\text{ROR}} + \tilde{u}_{\neg \text{CRIME}}^{\text{CASH}}$	$u_{\text{CASH}, \neg \text{CRIME}}$ $= u_{\neg \text{CRIME}}^{\text{CASH}} + \tilde{u}_{\neg \text{CRIME}}^{\text{ROR}}$
		Backlash ($Y(0) = 0, Y(1) = 1$)	$= u_{\neg \text{CRIME}}^{\text{ROR}} + \tilde{u}_{\text{CRIME}}^{\text{CASH}}$	$= u_{\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\neg \text{CRIME}}^{\text{ROR}}$
	$Y(0) = 1$	Preventable ($Y(0) = 1, Y(1) = 0$)	$u_{\text{ROR}, \text{CRIME}}$ $= u_{\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\neg \text{CRIME}}^{\text{CASH}}$	$u_{\text{CASH}, \text{CRIME}}$ $= u_{\neg \text{CRIME}}^{\text{CASH}} + \tilde{u}_{\text{CRIME}}^{\text{ROR}}$
		Hopeless ($Y(0) = 1, Y(1) = 1$)	$= u_{\text{CRIME}}^{\text{ROR}} + \tilde{u}_{\text{CRIME}}^{\text{CASH}}$	$= u_{\text{CRIME}}^{\text{CASH}} + \tilde{u}_{\text{CRIME}}^{\text{ROR}}$

- u_{d, y_0}

- $u_{0, y_0} = u_{y_0}^0$: standard utility under decision $d = 0$ with realized outcome $Y(0) = y_0$
- $u_{1, y_0} = \tilde{u}_{y_0}^1$: counterfactual utility under an alternative decision $d = 1$ with counterfactual outcome $Y(0) = y_0$

- risk score is a special case of triage score with restrictions

Nonparametric Identification of Expected Counterfactual Utility

- Expected counterfactual utility is generally unidentifiable

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

Nonparametric Identification of Expected Counterfactual Utility

- Expected counterfactual utility is generally unidentifiable

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

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

Nonparametric Identification of Expected Counterfactual Utility

- Expected counterfactual utility is generally unidentifiable

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

- Partial identification approaches (Li and Pearl, 2019; Ben-Michael et al. 2024)
- Additive counterfactual utility (Koch and Imai, 2025):

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

Nonparametric Identification of Expected Counterfactual Utility

- Expected counterfactual utility is generally unidentifiable

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

- Partial identification approaches (Li and Pearl, 2019; Ben-Michael et al. 2024)
- Additive counterfactual utility (Koch and Imai, 2025):

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

- standard utility as a special case

Nonparametric Identification of Expected Counterfactual Utility

- Expected counterfactual utility is generally unidentifiable

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

- Partial identification approaches (Li and Pearl, 2019; Ben-Michael et al. 2024)
- Additive counterfactual utility (Koch and Imai, 2025):

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

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

Nonparametric Identification of Expected Counterfactual Utility

- Expected counterfactual utility is generally unidentifiable

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

- Partial identification approaches (Li and Pearl, 2019; Ben-Michael et al. 2024)
- Additive counterfactual utility (Koch and Imai, 2025):

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

- standard utility as a special case
- can be a function of covariates $\mathbf{X}$
- Additivity as a necessary and sufficient identification condition under unconfoundedness

Nonparametric Identification of Expected Counterfactual Utility

- Expected counterfactual utility is generally unidentifiable

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

- Partial identification approaches (Li and Pearl, 2019; Ben-Michael et al. 2024)
- Additive counterfactual utility (Koch and Imai, 2025):

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

- standard utility as a special case
- can be a function of covariates $\mathbf{X}$
- Additivity as a **necessary and sufficient** identification condition under unconfoundedness
- Counterfactual utility may not have a standard utility representation beyond the binary decision case

Axiomatizing Counterfactual Utility (Koch, Imai, and Strzalecki, 2026)

- Recent criticisms about counterfactual utility

Axiomatizing Counterfactual Utility (Koch, Imai, and Strzalecki, 2026)

- Recent criticisms about counterfactual utility
 - Gelman and Mikhaeil (2025): counterfactual utility violates von Neumann-Morgenstern axioms

Axiomatizing Counterfactual Utility (Koch, Imai, and Strzalecki, 2026)

- Recent criticisms about counterfactual utility
 - Gelman and Mikhaeil (2025): counterfactual utility violates von Neumann-Morgenstern axioms
 - Sawant and Stensrud (2026): counterfactual utility can be intransitive

Axiomatizing Counterfactual Utility (Koch, Imai, and Strzalecki, 2026)

- Recent criticisms about counterfactual utility
 - Gelman and Mikhaeil (2025): counterfactual utility violates von Neumann-Morgenstern axioms
 - Sawant and Stensrud (2026): counterfactual utility can be intransitive
- Consider the **potential outcomes space** $\mathcal{Z} := \mathcal{D} \times \mathcal{Y}^{\mathcal{D}} \times \mathcal{X}$

Axiomatizing Counterfactual Utility (Koch, Imai, and Strzalecki, 2026)

- Recent criticisms about counterfactual utility
 - Gelman and Mikhaeil (2025): counterfactual utility violates von Neumann-Morgenstern axioms
 - Sawant and Stensrud (2026): counterfactual utility can be intransitive
- Consider the **potential outcomes space** $\mathcal{Z} := \mathcal{D} \times \mathcal{Y}^{\mathcal{D}} \times \mathcal{X}$
- Let $\Delta(\mathcal{Z})$ denote the set of probability distributions on $\mathcal{Z}$

Axiomatizing Counterfactual Utility (Koch, Imai, and Strzalecki, 2026)

- Recent criticisms about counterfactual utility
 - Gelman and Mikhaeil (2025): counterfactual utility violates von Neumann-Morgenstern axioms
 - Sawant and Stensrud (2026): counterfactual utility can be intransitive
- Consider the **potential outcomes space** $\mathcal{Z} := \mathcal{D} \times \mathcal{Y}^{\mathcal{D}} \times \mathcal{X}$
- Let $\Delta(\mathcal{Z})$ denote the set of probability distributions on $\mathcal{Z}$

Definition 1 (Preference Relation)

Let $P^\pi, Q^\rho \in \Delta(\mathcal{Z})$ denote the probability distributions induced by policies π and ρ under states of nature P and Q , respectively. Then, a binary preference relation $\succsim$ on $\Delta(\mathcal{Z})$ is represented by a counterfactual utility $\tilde{u}$ if the value of policy π in state P is no less than that of policy ρ in state Q , i.e.,

$$P^\pi \succsim Q^\rho \iff V_P(\pi; \tilde{u}) \geq V_Q(\rho; \tilde{u}).$$

where $V_P(\pi; \tilde{u}) := \mathbb{E}_{P^\pi}[\tilde{u}(D, Y(0), \dots, Y(K-1), \mathbf{X})]$

von Neumann-Morgenstern for Counterfactual Utility

- von Neumann-Morgenstern (vNM) axioms

von Neumann-Morgenstern for Counterfactual Utility

- von Neumann-Morgenstern (vNM) axioms
 - ① **Completeness**: For all $p, q \in \Delta(\mathcal{Z})$, either $p \succsim q$ or $q \succsim p$

von Neumann-Morgenstern for Counterfactual Utility

- von Neumann-Morgenstern (vNM) axioms
 - ① **Completeness**: For all $p, q \in \Delta(\mathcal{Z})$, either $p \succsim q$ or $q \succsim p$
 - ② **Transitivity**: For all $p, q, r \in \Delta(\mathcal{Z})$, if $p \succsim q$ and $q \succsim r$, then $p \succsim r$

von Neumann-Morgenstern for Counterfactual Utility

- von Neumann-Morgenstern (vNM) axioms
 - ① **Completeness**: For all $p, q \in \Delta(\mathcal{Z})$, either $p \succsim q$ or $q \succsim p$
 - ② **Transitivity**: For all $p, q, r \in \Delta(\mathcal{Z})$, if $p \succsim q$ and $q \succsim r$, then $p \succsim r$
 - ③ **Independence**: For all $p, q, r \in \Delta(\mathcal{Z})$ and $\alpha \in (0, 1]$, if $p \succ q$ then $\alpha p + (1 - \alpha)r \succ \alpha q + (1 - \alpha)r$

von Neumann-Morgenstern for Counterfactual Utility

- von Neumann-Morgenstern (vNM) axioms
 - ① **Completeness**: For all $p, q \in \Delta(\mathcal{Z})$, either $p \succsim q$ or $q \succsim p$
 - ② **Transitivity**: For all $p, q, r \in \Delta(\mathcal{Z})$, if $p \succsim q$ and $q \succsim r$, then $p \succsim r$
 - ③ **Independence**: For all $p, q, r \in \Delta(\mathcal{Z})$ and $\alpha \in (0, 1]$, if $p \succ q$ then $\alpha p + (1 - \alpha)r \succ \alpha q + (1 - \alpha)r$
 - ④ **Continuity**: For all $p, q, r \in \Delta(\mathcal{Z})$, if $p \succ q \succ r$ then there exists $\alpha, \beta \in (0, 1)$ such that $\alpha p + (1 - \alpha)r \succ q \succ \beta p + (1 - \beta)r$

von Neumann-Morgenstern for Counterfactual Utility

- von Neumann-Morgenstern (vNM) axioms
 - 1 **Completeness**: For all $p, q \in \Delta(\mathcal{Z})$, either $p \succsim q$ or $q \succsim p$
 - 2 **Transitivity**: For all $p, q, r \in \Delta(\mathcal{Z})$, if $p \succsim q$ and $q \succsim r$, then $p \succsim r$
 - 3 **Independence**: For all $p, q, r \in \Delta(\mathcal{Z})$ and $\alpha \in (0, 1]$, if $p \succ q$ then $\alpha p + (1 - \alpha)r \succ \alpha q + (1 - \alpha)r$
 - 4 **Continuity**: For all $p, q, r \in \Delta(\mathcal{Z})$, if $p \succ q \succ r$ then there exists $\alpha, \beta \in (0, 1)$ such that $\alpha p + (1 - \alpha)r \succ q \succ \beta p + (1 - \beta)r$

Theorem 1 (vNM for Counterfactual Utility)

The following statements are equivalent:

- 1 $\succsim$ satisfies the above axioms;
- 2 there exists a counterfactual utility $\tilde{u} : \mathcal{Z} \rightarrow \mathbb{R}$ that represents $\succsim$ in the sense of Definition ??.

Moreover, any such utility function $\tilde{u}$ is unique up to a positive affine transformation.

von Neumann-Morgenstern for Counterfactual Utility

- von Neumann-Morgenstern (vNM) axioms

- 1 **Completeness:** For all $p, q \in \Delta(\mathcal{Z})$, either $p \succsim q$ or $q \succsim p$
- 2 **Transitivity:** For all $p, q, r \in \Delta(\mathcal{Z})$, if $p \succsim q$ and $q \succsim r$, then $p \succsim r$
- 3 **Independence:** For all $p, q, r \in \Delta(\mathcal{Z})$ and $\alpha \in (0, 1]$, if $p \succ q$ then $\alpha p + (1 - \alpha)r \succ \alpha q + (1 - \alpha)r$
- 4 **Continuity:** For all $p, q, r \in \Delta(\mathcal{Z})$, if $p \succ q \succ r$ then there exists $\alpha, \beta \in (0, 1)$ such that $\alpha p + (1 - \alpha)r \succ q \succ \beta p + (1 - \beta)r$

Theorem 1 (vNM for Counterfactual Utility)

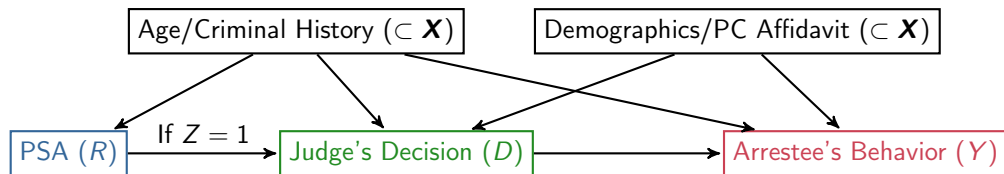
The following statements are equivalent:

- 1 $\succsim$ satisfies the above axioms;
- 2 there exists a counterfactual utility $\tilde{u} : \mathcal{Z} \rightarrow \mathbb{R}$ that represents $\succsim$ in the sense of Definition ??.

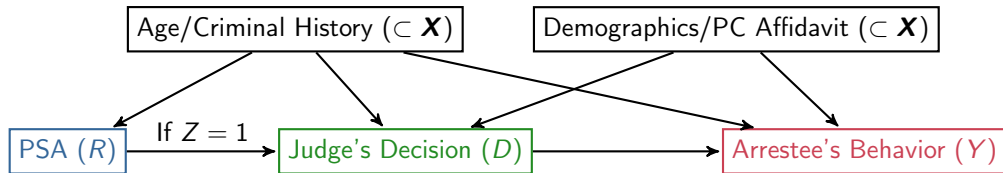
Moreover, any such utility function $\tilde{u}$ is unique up to a positive affine transformation.

- Key point: one must evaluate counterfactual utility in the potential outcomes space

Assumptions in Our Empirical Application

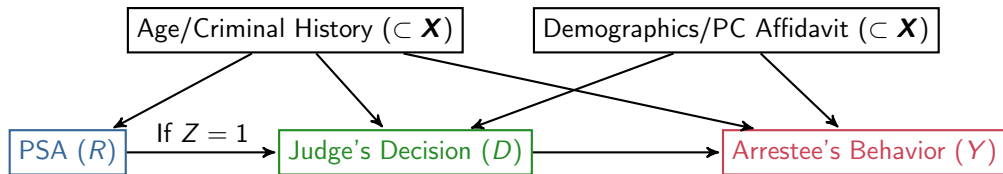


Assumptions in Our Empirical Application



- Assumption 1: Single-blinded and unconfounded treatment assignment

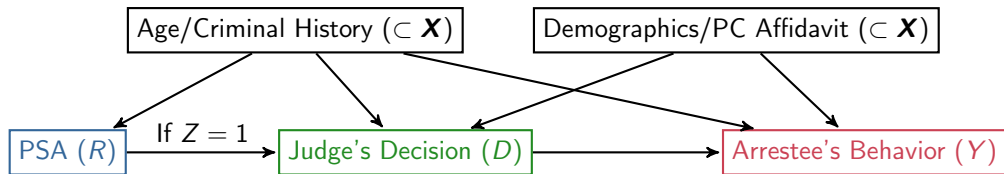
Assumptions in Our Empirical Application



- Assumption 1: **Single-blinded and unconfounded treatment assignment**

- ① Single-blinded: $Y_i(Z_i = 0, D_i(Z_i) = d) = Y_i(Z_i = 1, D_i(Z_i) = d) = Y_i(d)$ for $d = 0, 1$
 $\leadsto$ PSA can only affect the **outcome** through the **decision**

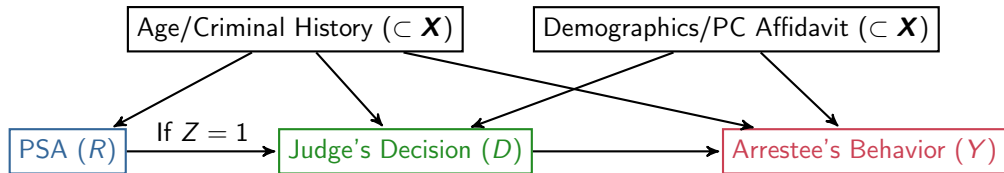
Assumptions in Our Empirical Application



- Assumption 1: **Single-blinded and unconfounded treatment assignment**

- 1 Single-blinded: $Y_i(Z_i = 0, D_i(Z_i) = d) = Y_i(Z_i = 1, D_i(Z_i) = d) = Y_i(d)$ for $d = 0, 1$
 $\leadsto$ PSA can only affect the **outcome** through the **decision**
- 2 Unconfounded: $Z_i \perp\!\!\!\perp \{R_i, \{D_i(z), Y_i(d)\}_{z \in \{0,1\}, d \in \mathcal{D}}\} \mid \mathbf{X}_i$

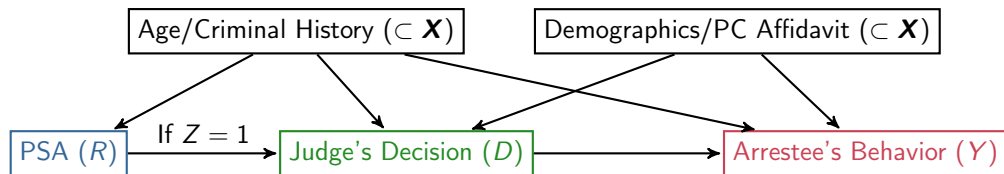
Assumptions in Our Empirical Application



- Assumption 1: **Single-blinded and unconfounded treatment assignment**

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

Assumptions in Our Empirical Application



- Assumption 1: **Single-blinded and unconfounded treatment assignment**

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

- Assumption 2: **Unconfounded decision** and overlap

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

This holds so long as we have all the information that the judge takes into account

Affidavit of Probable Cause and GenAI Powered Inference (GPI)

Affidavit of Probable Cause and GenAI Powered Inference (GPI)

IN THE
[REDACTED] JUSTICE COURT
COUNTY OF [REDACTED], STATE OF UTAH

State of Utah

vs.

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] 2020 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: [REDACTED]

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]	SIU: [REDACTED] UIN: [REDACTED]
Booking UserID: [REDACTED]	Booking date/time: [REDACTED] Submission ID: [REDACTED] (Version 1)

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

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

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] 2020 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: [REDACTED]

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]	SIU: [REDACTED] UIN: [REDACTED]
Booking UserID: [REDACTED]	Booking date/time: [REDACTED] Submission ID: [REDACTED] (Version 1)

Affidavit of Probable Cause and GenAI Powered Inference (GPI)

IN THE
[REDACTED] JUSTICE COURT
COUNTY OF [REDACTED], STATE OF UTAH

State of Utah

vs.

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] 2020 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: [REDACTED]

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]	SIU: [REDACTED] UIN: [REDACTED]
Booking User ID: [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

Affidavit of Probable Cause and GenAI Powered Inference (GPI)

IN THE
[REDACTED] JUSTICE COURT
COUNTY OF [REDACTED] STATE OF UTAH

State of Utah

vs.

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: [REDACTED]

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]	SIU: [REDACTED] UIN: [REDACTED]
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 2026)
 - use open-source LLM to regenerate texts
 - obtain their internal representation

Affidavit of Probable Cause and GenAI Powered Inference (GPI)

IN THE
[REDACTED] JUSTICE COURT
COUNTY OF [REDACTED] STATE OF UTAH

State of Utah

vs.

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] 2020 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: [REDACTED]

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]	SIU: [REDACTED] UIN: [REDACTED]
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 2026)
 - use open-source LLM to regenerate texts
 - obtain their internal representation
 - guaranteed to contain all the information

Affidavit of Probable Cause and GenAI Powered Inference (GPI)

IN THE
[REDACTED] JUSTICE COURT
COUNTY OF [REDACTED] STATE OF UTAH

State of Utah	Affidavit of Probable Cause
vs. [REDACTED] Date of Birth: [REDACTED]	
Arrestee	

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: [REDACTED]

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]	SIU: [REDACTED] UIN: [REDACTED]
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 2026)
 - use open-source LLM to regenerate texts
 - obtain their internal representation
 - guaranteed to contain all the information
 - leverage LLM's internal structure

Policy Evaluation and Learning

- Nonparametric identification of expected additive counterfactual utilities

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

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

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

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

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
 - semiparametric inference: asymptotic normality, double-robustness, rate-robustness

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
 - semiparametric inference: asymptotic normality, double-robustness, rate-robustness
- **Statistical hypothesis test**

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
 - semiparametric inference: asymptotic normality, double-robustness, rate-robustness
- **Statistical hypothesis test**
 - ① **When do we prefer Human+PSA?** Invert the hypothesis test

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

for a given counterfactual utility function $\tilde{u}$

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
 - semiparametric inference: asymptotic normality, double-robustness, rate-robustness
- **Statistical hypothesis test**
 - ① **When do we prefer Human+PSA?** Invert the hypothesis test

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

for a given counterfactual utility function $\tilde{u}$

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

Utility Specification

- Focus on **cost** and **regret** (they are all non-negative):

Utility Specification

- Focus on **cost** and **regret** (they are all non-negative):
 - Cost of decision: c^{CASH}

Utility Specification

- Focus on **cost and regret** (they are all non-negative):
 - Cost of decision: c^{CASH}
 - Cost of outcome: $c_{\text{CRIME}}^{\text{ROR}}$ and $c_{\text{CRIME}}^{\text{CASH}}$

Utility Specification

- Focus on **cost and regret** (they are all non-negative):
 - Cost of decision: c^{CASH}
 - Cost of outcome: $c_{\text{CRIME}}^{\text{ROR}}$ and $c_{\text{CRIME}}^{\text{CASH}}$
 - Regret from counterfactual outcome: $r_{\neg\text{CRIME}}^{\text{ROR}}$ and $r_{\neg\text{CRIME}}^{\text{CASH}}$

Utility Specification

- Focus on **cost** and **regret** (they are all non-negative):
 - Cost of decision: c^{CASH}
 - Cost of outcome: $c_{\text{CRIME}}^{\text{ROR}}$ and $c_{\text{CRIME}}^{\text{CASH}}$
 - Regret from counterfactual outcome: $r_{\neg\text{CRIME}}^{\text{ROR}}$ and $r_{\neg\text{CRIME}}^{\text{CASH}}$

Decision

		Release ($D^* = 0$)	Cash bail ($D^* = 1$)
Principal Strata	Safe ($Y(0) = 0, Y(1) = 0$)	$-r_{\neg\text{CRIME}}^{\text{CASH}}$	$-c^{\text{CASH}} - r_{\neg\text{CRIME}}^{\text{ROR}}$
	Backlash ($Y(0) = 0, Y(1) = 1$)	0	$-c^{\text{CASH}} - c_{\text{CRIME}}^{\text{CASH}} - r_{\neg\text{CRIME}}^{\text{ROR}}$
	Preventable ($Y(0) = 1, Y(1) = 0$)	$-c_{\text{CRIME}}^{\text{ROR}} - r_{\neg\text{CRIME}}^{\text{CASH}}$	$-c^{\text{CASH}}$
	Hopeless ($Y(0) = 1, Y(1) = 1$)	$-c_{\text{CRIME}}^{\text{ROR}}$	$-c^{\text{CASH}} - c_{\text{CRIME}}^{\text{CASH}}$

Utility Specification (cont.)

- Standardize $c_{\text{CRIME}}^{\text{ROR}} = 1$ (cost of new crime under ROR)

Utility Specification (cont.)

- Standardize $c_{\text{CRIME}}^{\text{ROR}} = 1$ (cost of new crime under ROR)
 - ① **standard** decision theory with the **risk** score framework: c^{CASH}

Utility Specification (cont.)

- Standardize $c_{\text{CRIME}}^{\text{ROR}} = 1$ (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} , $c_{\text{CRIME}}^{\text{CASH}}$

Utility Specification (cont.)

- Standardize $c_{\text{CRIME}}^{\text{ROR}} = 1$ (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: $c^{\text{CASH}}, c_{\text{CRIME}}^{\text{CASH}}$
 - ③ **counterfactual** decision theory with the **risk** score framework: $c^{\text{CASH}}, r_{\neg\text{CRIME}}^{\text{ROR}}$

Utility Specification (cont.)

- Standardize $c_{\text{CRIME}}^{\text{ROR}} = 1$ (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: $c^{\text{CASH}}, c_{\text{CRIME}}^{\text{CASH}}$
 - ③ **counterfactual** decision theory with the **risk** score framework: $c^{\text{CASH}}, r_{\neg\text{CRIME}}^{\text{ROR}}$
 - ④ **counterfactual** decision theory with the **triage** score framework: $c^{\text{CASH}}, c_{\text{CRIME}}^{\text{CASH}}, r_{\neg\text{CRIME}}^{\text{ROR}}, r_{\neg\text{CRIME}}^{\text{CASH}}$

Utility Specification (cont.)

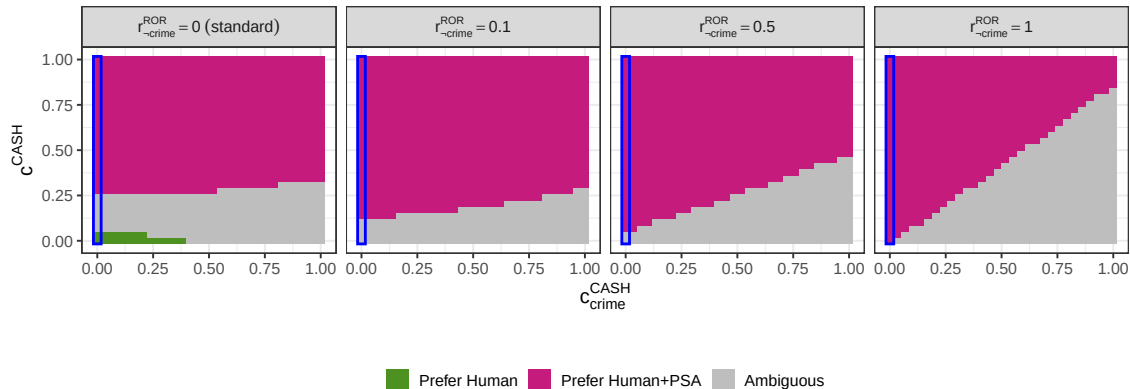
- Standardize $c_{\text{CRIME}}^{\text{ROR}} = 1$ (cost of new crime under ROR)
 - ① **standard** decision theory with the **risk** score framework: c^{CASH}
 - ② **standard** decision theory with the **triage** score framework: $c^{\text{CASH}}, c_{\text{CRIME}}^{\text{CASH}}$
 - ③ **counterfactual** decision theory with the **risk** score framework: $c^{\text{CASH}}, r_{\neg\text{CRIME}}^{\text{ROR}}$
 - ④ **counterfactual** decision theory with the **triage** score framework: $c^{\text{CASH}}, c_{\text{CRIME}}^{\text{CASH}}, r_{\neg\text{CRIME}}^{\text{ROR}}, r_{\neg\text{CRIME}}^{\text{CASH}}$
- To further reduce the number of parameters for our analysis:

$$\begin{aligned}
 & \frac{\text{Cost of new crime under cash bail } (c_{\text{CRIME}}^{\text{CASH}})}{\text{Cost of new crime under ROR } (c_{\text{CRIME}}^{\text{ROR}})} \\
 &= \frac{\text{Regret of no new crime under cash bail } (r_{\neg\text{CRIME}}^{\text{CASH}})}{\text{Regret of no new crime under ROR } (r_{\neg\text{CRIME}}^{\text{ROR}})}
 \end{aligned}$$

cost ratio is the same as regret ratio $\leadsto r_{\neg\text{CRIME}}^{\text{CASH}} = r_{\neg\text{CRIME}}^{\text{ROR}} c_{\text{CRIME}}^{\text{CASH}}$

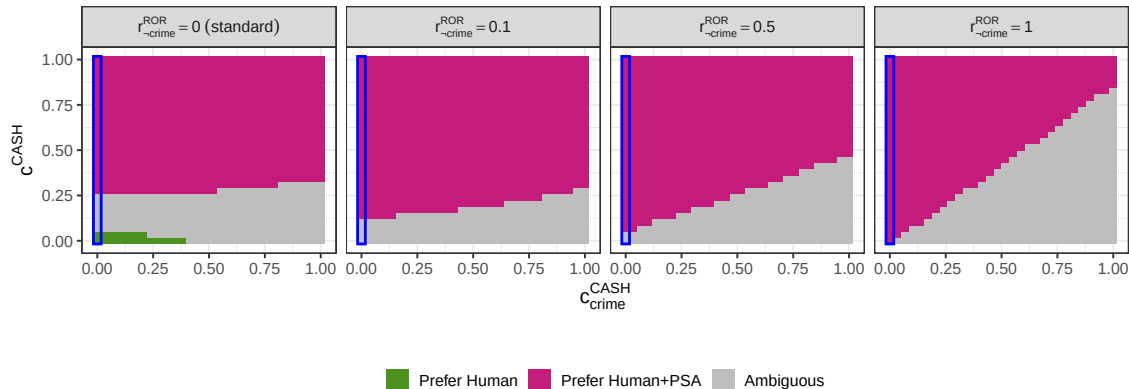
When Do We Prefer Human+PSA over Human?

When Do We Prefer Human+PSA over Human?



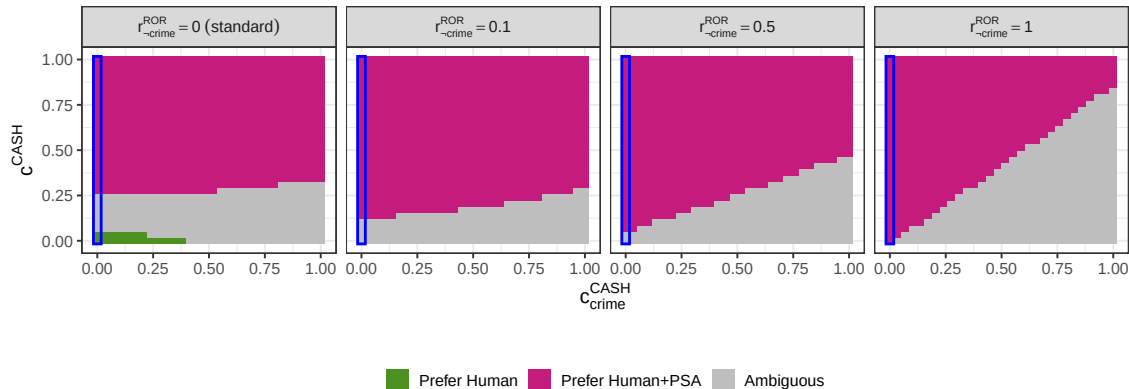
- Outcome: new criminal activity

When Do We Prefer Human+PSA over Human?



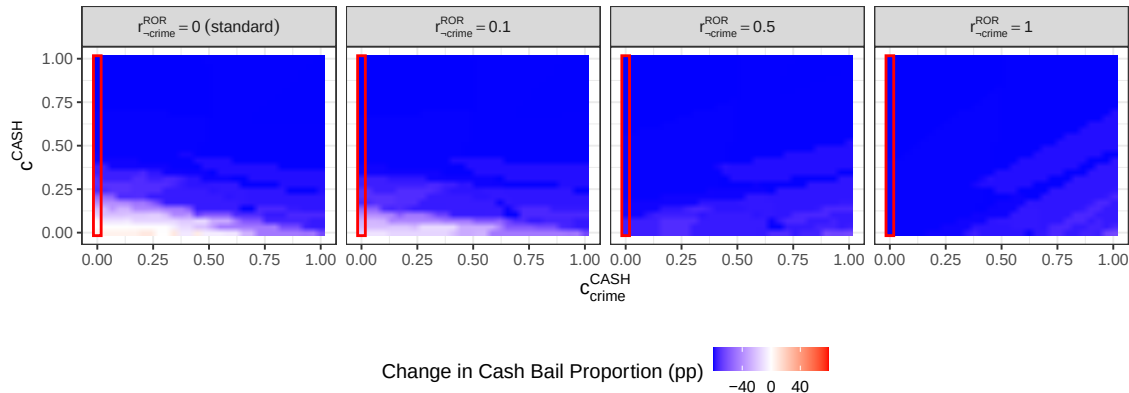
- Outcome: new criminal activity
- Blue boxes: risk score system (when $c_{CRIME}^{CASH} = 0$)

When Do We Prefer Human+PSA over Human?



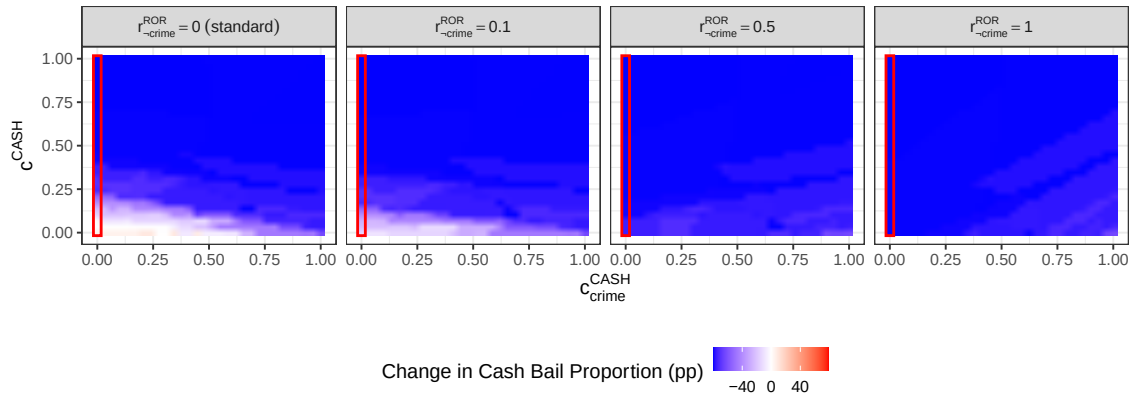
- Outcome: new criminal activity
- Blue boxes: risk score system (when $c_{\text{CRIME}}^{\text{CASH}} = 0$)
- First panel: standard decision framework (when $r_{\neg\text{CRIME}}^{\text{ROR}} = r_{\text{CRIME}}^{\text{ROR}} = 0$)
- More ambiguous under triage score than risk score

Optimal Decision Rules



- Outcome: new criminal activity
- Policy class: decision trees (depth 2; minimum leaf size 10)
- Most utility settings favor less cash bail than the current practice (blue)

Optimal Decision Rules



- Outcome: new criminal activity
- Policy class: decision trees (depth 2; minimum leaf size 10)
- Most utility settings favor less cash bail than the current practice (blue)
- Triage and risk scores can imply different cash-bail policies

Concluding Remarks

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

Concluding Remarks

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

- Unlike risk score, triage score can incorporate counterfactual utilities

Concluding Remarks

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

- Unlike risk score, triage score can incorporate counterfactual utilities
- It enables decision-makers to incorporate a wide range of ethical and practical factors

Concluding Remarks

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

- Unlike risk score, triage score can incorporate counterfactual utilities
- It enables decision-makers to incorporate a wide range of ethical and practical factors
- Triage score captures richer utility structures than risk score and yield distinct results

Concluding Remarks

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

- Unlike risk score, triage score can incorporate counterfactual utilities
- It enables decision-makers to incorporate a wide range of ethical and practical factors
- Triage score captures richer utility structures than risk score and yield distinct results
- We apply our framework to a RCT to evaluate **a pre-trial risk assessment instrument**

Concluding Remarks

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

- Unlike risk score, triage score can incorporate counterfactual utilities
- It enables decision-makers to incorporate a wide range of ethical and practical factors
- Triage score captures richer utility structures than risk score and yield distinct results
- We apply our framework to a RCT to evaluate **a pre-trial risk assessment instrument**
- The triage score framework yields higher expected utility for Human+PSA and more lenient optimal recommendations than the risk score framework

Concluding Remarks

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

- Unlike risk score, triage score can incorporate counterfactual utilities
- It enables decision-makers to incorporate a wide range of ethical and practical factors
- Triage score captures richer utility structures than risk score and yield distinct results
- We apply our framework to a RCT to evaluate **a pre-trial risk assessment instrument**
- The triage score framework yields higher expected utility for Human+PSA and more lenient optimal recommendations than the risk score framework
- Next steps:

Concluding Remarks

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

- Unlike risk score, triage score can incorporate counterfactual utilities
- It enables decision-makers to incorporate a wide range of ethical and practical factors
- Triage score captures richer utility structures than risk score and yield distinct results
- We apply our framework to a RCT to evaluate **a pre-trial risk assessment instrument**
- The triage score framework yields higher expected utility for Human+PSA and more lenient optimal recommendations than the risk score framework
- Next steps:
 - Ongoing partnerships with courts, prosecutors' offices, and child protection agencies

Concluding Remarks

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

- Unlike risk score, triage score can incorporate counterfactual utilities
- It enables decision-makers to incorporate a wide range of ethical and practical factors
- Triage score captures richer utility structures than risk score and yield distinct results
- We apply our framework to a RCT to evaluate **a pre-trial risk assessment instrument**
- The triage score framework yields higher expected utility for Human+PSA and more lenient optimal recommendations than the risk score framework
- Next steps:
 - Ongoing partnerships with courts, prosecutors' offices, and child protection agencies
 - Evaluate Human-alone vs. Human+Triage decision systems