An Axiomatic Foundation for Decisions with Counterfactual Utility

Benedikt Koch, Kosuke Imai and Tomasz Strzalecki

Harvard University

ACIC 2026

Overview

Decision making with counterfactual utility:

- **Appeal**: encodes asymmetric criteria such as harm and regret;
- **Concern**: criticized as incoherent and intransitive.

What we do:

- Introduce potential outcome space
- Show counterfactual utility satisfies vNM axioms on this space
- Contrast with claims on the realized outcome space
- Investigate findings by Gelman and Mikhaeil 2025

Statistical Decision Theory

Wald 1950: Decision-making as a game against nature.

1. Nature picks an unknown state θ,
2. Decision-maker chooses action $D = d$,
3. A **loss** $\ell(d, \theta)$ quantifies choosing d under θ.

Given covariates $\boldsymbol{X}$, construct a decision rule $D = \pi(\boldsymbol{X})$.

Measure performance with **risk**,

$$R_\theta(\pi; \ell) = \mathbb{E}_\theta \left[\ell(\pi(\boldsymbol{X}), \theta) \right].$$

Foundation of estimation and hypothesis testing.

Standard Decision Theory for Treatment Choice

Manski 2004: Choose treatment D based on outcome $Y(D)$.

1. Nature picks an unknown state $\theta = P(Y(0), \ldots, Y(K-1), \boldsymbol{X})$,
2. Decision-maker chooses treatment $D = d$,
3. A **utility** $u(d, y_d)$ quantifies choosing d for outcome y_d.

Given covariates $\boldsymbol{X}$, construct a decision rule $D = \pi(\boldsymbol{X})$.

Measure performance with **value function**,

$$V_P(\pi; u) = \mathbb{E}_P\left[u(\pi(\boldsymbol{X}), Y(\pi(\boldsymbol{X})))\right].$$

Limitation: Utility depends only on *realized* outcome

Counterfactual Decision Theory

What happens if

$$V_P(\pi; \tilde{u}) = \mathbb{E}_P \left[\tilde{u}(\pi(\boldsymbol{X}), Y(0), \ldots, Y(K-1))\right]?$$

Raises two questions:

1. **Identification:** Only observe one potential outcome (Koch and Imai 2025)
2. **Coherence:** Is counterfactual decision making rational?

Allais paradox I (Kahneman and Tversky 1979)

Experiment 1:

- **A:** \$4000 with probability 0.8
- **B:** \$3000 for sure

Experiment 2:

- **C:** \$4000 with probability 0.2
- **D:** \$3000 with probability 0.25

Typically: $B \succ A$ and $C \succ D$.

Allais paradox II

Let $u(d; y_d) = u(y_d)$

$$B \succ A \Rightarrow u(3000) > 0.8\, u(4000)$$

$$C \succ D \Rightarrow u(3000) < 0.8\, u(4000)$$

Contradiction! Standard utilities can't explain this!

Allais paradox III

Resolve with counterfactual decision making:

$$\tilde{u}(d; y_0, y_1) = y_d + f(y_d - y_{1-d})$$

Condition	Interpretation	Implication
$y_d > y_{1-d}$	Rejoicing	$f(y_d - y_{1-d}) > 0$
$y_d < y_{1-d}$	Regret	$f(y_d - y_{1-d}) < 0$

If Regret $>$ Rejoicing, then

$$\underbrace{B \succ A}_{\text{Experiment 1}} \succ \underbrace{C \succ D}_{\text{Experiment 2}}$$

Preferences on Potential Outcome Space

Consider potential outcome space

$$(D, Y(0), \ldots, Y(K-1), \boldsymbol{X}) \in \mathcal{D} \times \mathcal{Y}^{\mathcal{D}} \times \mathcal{X} =: \mathcal{Z}$$

- Policy: $\pi : \mathcal{Y}^{\mathcal{D}} \times \mathcal{X} \to \Delta(\mathcal{D})$
- State of Nature: $P \in \Delta(\mathcal{Y}^{\mathcal{D}} \times \mathcal{X})$
- Induced Law: $P^{\pi}(d, \boldsymbol{y}, \boldsymbol{x}) = \pi(d \mid \boldsymbol{y}, \boldsymbol{x}) \cdot P(\boldsymbol{y}, \boldsymbol{x}) \in \Delta(\mathcal{Z})$

Counterfactual utility $\tilde{u} : \mathcal{Z} \to \mathbb{R}$ induces preferences on $\Delta(\mathcal{Z})$:

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

von Neumann-Morgenstern Axioms

Let $p, q, r \in \Delta(\mathcal{Z})$ and $\succsim$ a generic preference relation.

Axiom (Completeness)

Either $p \succsim q$ or $q \succsim p$.

Axiom (Transitivity)

If $p \succsim q$ and $q \succsim r$, then $p \succsim r$.

Axiom (Independence, informal)

Mixing two options with the same third option preserves ranking

Axiom (Continuity, informal)

Preferences have no discontinuous jumps

von Neumann–Morgenstern Theorem on $\Delta(\mathcal{Z})$

Theorem (K, Imai, Strzalecki)

The following are equivalent:

1. $\succsim$ *satisfies the vNM axioms on* $\Delta(\mathcal{Z})$*;*
2. *there exists* $\tilde{u} : \mathcal{Z} \to \mathbb{R}$ *such that* $\succsim = \succsim_{\tilde{u}}$.

Moreover, $\tilde{u}$ *is unique up to a positive affine transformation.*

Key Distinction:

1. Counterfactual decision making is coherent if you care about counterfactuals
2. But may not be if you only care about realized outcomes $Y = Y(D)$

Main Message

Decision making with both standard and counterfactual utilities is rational; it depends on your preferences.

Thank you!

benediktjkoch.github.io

Identification

arXiv:2505.08908

Coherence

arXiv:2605.05521

References I

- Gelman, Andrew and Jonas M Mikhaeil (2025). "Russian roulette: the need for stochastic potential outcomes when utilities depend on counterfactuals". In: *Biometrika* 112.4. ISSN: 1464-3510. DOI: 10.1093/biomet/asaf062. URL: http://dx.doi.org/10.1093/biomet/asaf062.
- Kahneman, Daniel and Amos Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk". In: *Econometrica* 47.2, pp. 263–291. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1914185 (visited on 04/19/2025).
- Koch, Benedikt and Kosuke Imai (2025). *Statistical Decision Theory with Counterfactual Loss*. arXiv: 2505.08908 [math.ST]. URL: https://arxiv.org/abs/2505.08908.
- Manski, Charles F (2004). "Statistical treatment rules for heterogeneous populations". In: *Econometrica* 72.4, pp. 1221–1246.

References II

Wald, Abraham (1950). *Statistical decision functions*. Wiley.

Russian Roulette (Gelman and Mikhaeil 2025)

Let $Y = 1$ denote survival and $Y = 0$ death.

Gun	Survival probability	$V_P(d; u)$	$V_P(d; \tilde{u}_{\mathrm{GM}})$
$d = 0$	$P(Y(0) = 1) = \frac{5}{6}$	$\frac{5}{6}$	$\frac{5}{42}$
$d = 1$	$P(Y(1) = 1) = \frac{6}{7}$	$\frac{6}{7}$	$\frac{3}{42}$

$$u(d; y_d) = y_d$$

$$\tilde{u}_{\mathrm{GM}}(d; y_0, y_1) = \begin{cases} \mathbb{1}\{y_0 > y_1\}, & d = 0, \\ \frac{1}{2}\mathbb{1}\{y_0 < y_1\}, & d = 1. \end{cases}$$

Claim: Counterfactual utility can give nonsensical decision recommendations

Utilities are Subjective and Context-Dependent

Let $Y = 1$ denote survival and $Y = 0$ death.

Gun	Survival probability	$V_P(d;u)$	$V_P(d;\tilde{u}_{\mathrm{GM}})$	$V_P(d;u_D)$	$V_P(d;u_C)$
$d=0$	$P(Y(0)=1)=\frac{5}{6}$	–	✗	✗	✗
$d=1$	$P(Y(1)=1)=\frac{6}{7}$	✓	–	–	–

$$u(d;y_d) = y_d$$

$$u_D(d;y_d) = -y_d$$

$$\tilde{u}_{\mathrm{GM}}(d;y_0,y_1) = \begin{cases} \mathbb{1}\{y_0 > y_1\}, & d=0, \\ \frac{1}{2}\mathbb{1}\{y_0 < y_1\}, & d=1 \end{cases}$$

$$u_C(d;y_d) = y_d - c \cdot d$$

Russian Roulette in the Medical Context

Medical context

- $d = 0$: standard care
- $d = 1$: experimental treatment
- $Y = 1$: survival, $Y = 0$: death

$$\mathbb{E}[Y(0)] \approx 83\%, \qquad \mathbb{E}[Y(1)] \approx 86\%$$

$$\text{Benefit} : P(Y(0) = 0, Y(1) = 1) \approx 14\%, \quad \text{Harm} : P(Y(0) = 1, Y(1) = 0) \approx 12\%.$$

Interpretation: Counterfactual utility prevents harm

Classical vs. Counterfactual Decision Theory

	Classical	**Counterfactual**
Primary Object	$Y = Y(D)$	$(D, Y(0), \ldots, Y(K-1))$
Preferences	$\Delta(\mathcal{Y})$	$\Delta(\mathcal{D} \times \mathcal{Y}^{\mathcal{D}})$

Key distinction

Counterfactual preferences are transitive on $\Delta(\mathcal{D} \times \mathcal{Y}^{\mathcal{D}})$ but may not be on $\Delta(\mathcal{Y})$.

Projecting Back to Realized Outcomes

A counterfactual preference on $\Delta(\mathcal{Z})$ can induce choices among realized-outcome lotteries.

- **Menu-dependent projection:** compare lotteries using only the menu at hand.

$$\chi^{\mathrm{M}}(\mathcal{A}) = \operatorname*{argmax}_{d\in\mathcal{A}} \mathbb{E}_{P_{\mathcal{A}}}[\tilde{u}_{\mathcal{A}}(d;(Y(k))_{k\in\mathcal{A}})] \,.$$

May induce intransitive revealed preferences.

- **Context-dependent projection:** fix the full potential-outcome context first.

$$\chi^{\mathrm{C}}_{(\tilde{u},P)}(\mathcal{A}) = \operatorname*{argmax}_{d\in\mathcal{A}} \mathbb{E}_{P}[\tilde{u}(d;Y(0),\ldots,Y(K-1))].$$

Complete and transitive, but context-dependent.