LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

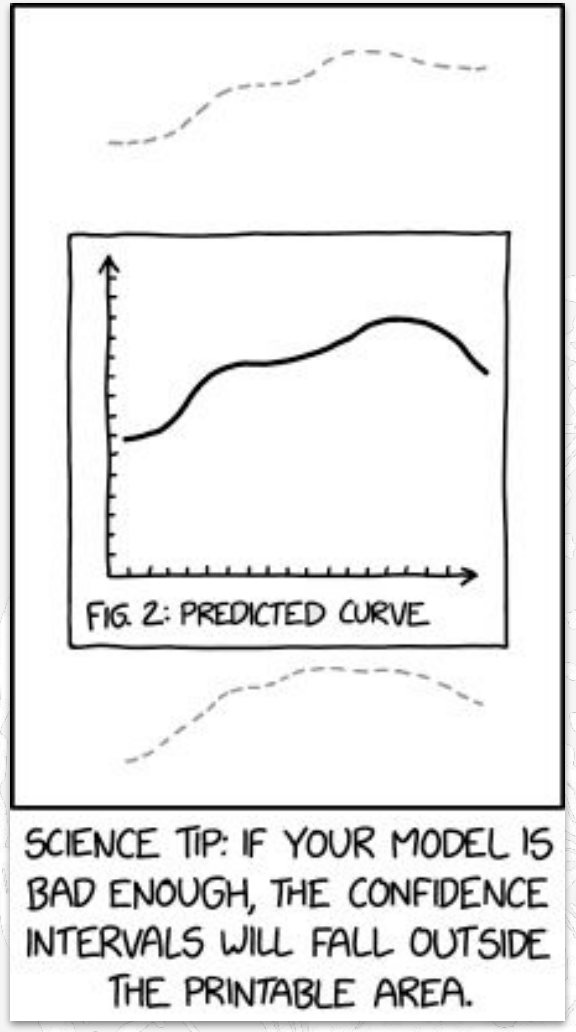
INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

mcml Munich Center for Machine Learning

Uncertainty quantification in Causal ML

Valentyn Melnychuk
LMU Munich

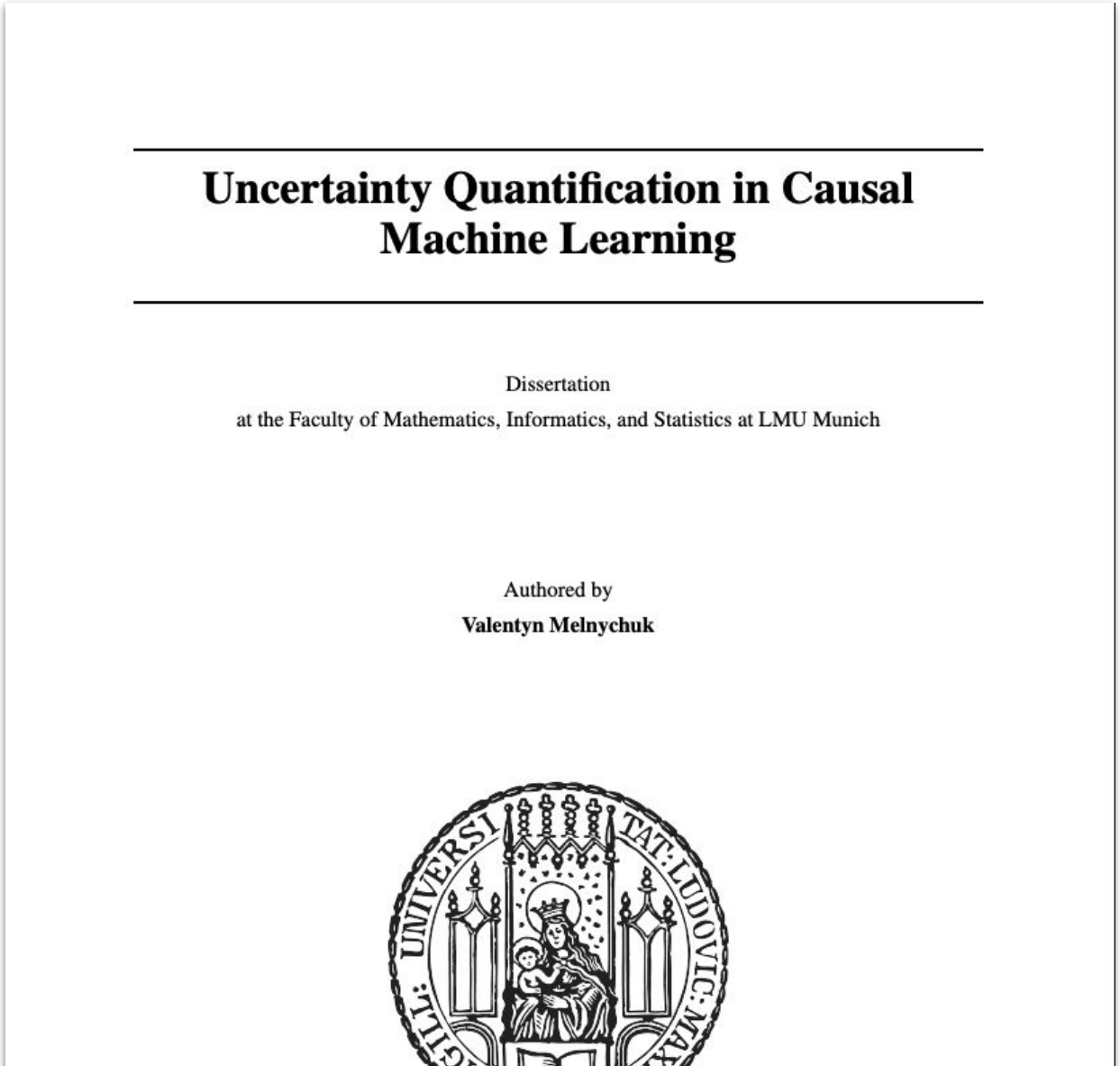
5th MCML Workshop on Causal Machine Learning
13.08.2026 | 10:10 - 11:10



https://www.explainxkcd.com/wiki/index.php/2311:_Confidence_Interval

Reference

- This presentation is based on the introduction my dissertation
- Reference:
  Melnychuk, V., Uncertainty Quantification in Causal Machine Learning, LMU Munich, 2026 (**will be published soon**)
- Today, I aim to provide a structured tour over uncertainty quantification methods in causal ML



Uncertainty Quantification in Causal Machine Learning

Dissertation

at the Faculty of Mathematics, Informatics, and Statistics at LMU Munich

Authored by

Valentyn Melnychuk

Agenda

Introduction

Epistemic uncertainty

Aleatoric uncertainty

Uncertainty due to unobserved variables

Combining uncertainties

Takeaways

Agenda

Introduction:

- Causal ML & Uncertainty
- Problem formulation
- Causal assumptions
- Spectrum of causal estimands
- Can we avoid causal assumptions?
- Three sources of uncertainty
- (Non-exhaustive) overview of approaches

Epistemic uncertainty

Aleatoric uncertainty

Uncertainty due to unobserved variables

Combining uncertainties

Takeaways

Introduction: Causal ML & Uncertainty

What is causal ML?

- **Causal ML** bridges structural causal reasoning (Rubin, 1974; Pearl, 2009; Bareinboim et al., 2022 [1]) and flexible predictive models
- Prediction targets are inherently causal:
  - potential outcomes (POs) $Y[a]$
  - treatment effects (TEs) $Y[a] - Y[a']$
- Applications: personalized medicine (Feuerriegel et al., 2024 [2]), public policy, marketing, education — wherever RCTs are limited or infeasible
  - Studies on health impact of smoking and alcohol consumption (Cornfield et al., 1959; Griswold et al. 2018 [3]) are possible due to 20th-century advances in causal inference

But: in sensitive domains, a point estimate is not enough — reliable deployment requires **quantifying the uncertainty** around it

[1] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5), 688–701. · Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. · Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 507–556. ACM.

[2] Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I. S., & van der Schaar, M. (2024). Causal machine learning for predicting treatment outcomes. *Nature Medicine* 30(4), 958–968. · Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. Applied causal inference powered by ML and AI. arXiv:2403.02467.

[3] Cornfield, J. et al. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *JNCI* 22(1), 173–203. · Griswold, M. G. et al. (2018). Alcohol use and burden for 195 countries and territories, 1990–2016. *The Lancet* 392(10152), 1015–1035.

Introduction: Problem formulation

Problem formulation

Given i.i.d. observational dataset $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^n \sim \mathbb{P}(X, A, Y)$ with

- X covariates
- A binary treatments
- Y continuous (factual) outcomes

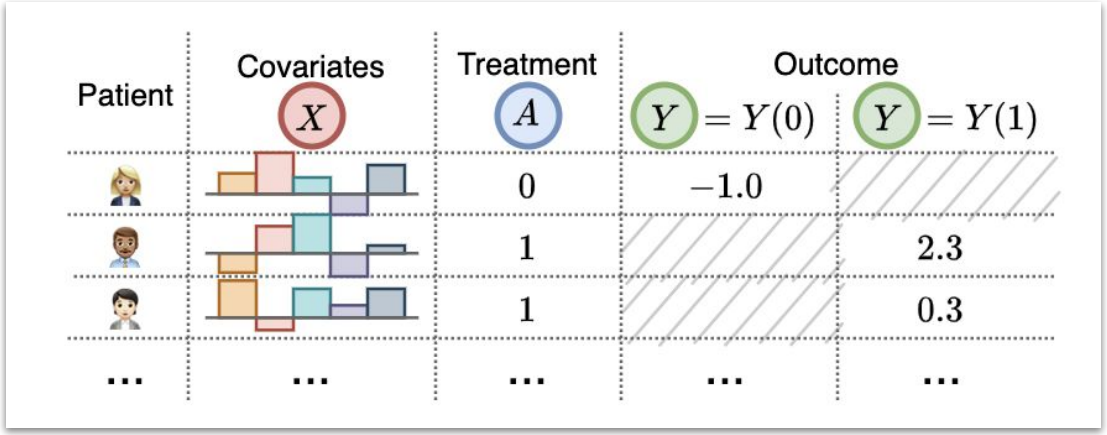
Given covariates, causal ML aims to predict causal targets:

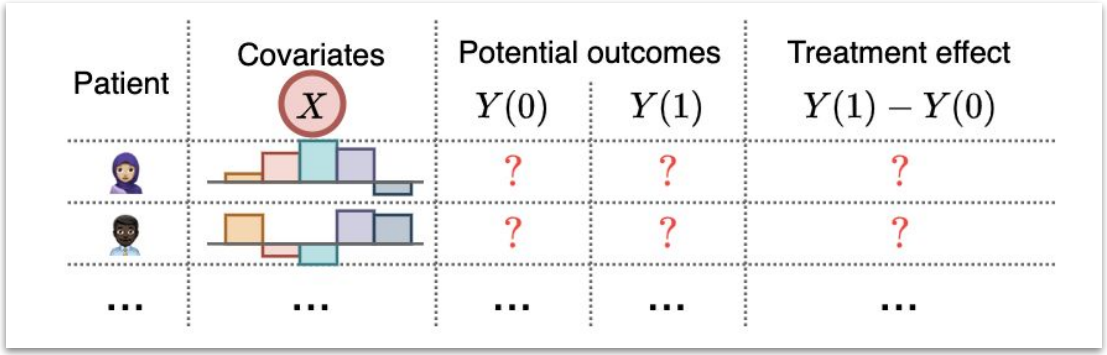
- potential outcomes: $Y[a]$
- treatment effects: $Y[a] - Y[a']$

However, we **never** observe both potential (counterfactual) outcomes!
→ we face a missing data problem

I use the following **nuisance functions**:

- propensity score $\pi_a(x) = \mathbb{P}(A = a \mid X = x)$
- outcome regression $\mu_a(x) = \mathbb{E}(Y \mid x, a)$
- conditional CDF $F_a(y \mid x) = \mathbb{P}(Y \leq y \mid x, a)$





Introduction: Causal assumptions

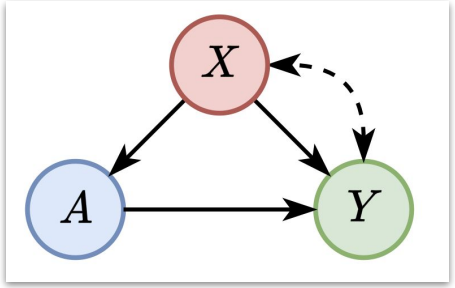
Standard identifiability assumptions

- Potential outcomes framework (Neyman-Rubin):

  1. Consistency. If $A = a$, then $Y = Y[a]$

  2. Strong overlap. There is always a non-zero probability of receiving treatment, conditioning on the covariates: $\epsilon > 0, \mathbb{P}(1 - \epsilon \geq \pi_a(X) \geq \epsilon) = 1$

  3. Unconfoundedness. Current treatment is independent of the potential outcome, conditioning on the covariates: $A \perp\!\!\!\perp Y[a] \mid X$ for all a.



The causal diagram of a DGP that satisfies Assumptions (1) - (3)

Additional assumptions

- Later, I also additionally assume a bijective generative mechanism (BGM) / rank invariance assumption for Y:
  - $\mathbb{F}_a(Y[a] \mid X) = \mathbb{F}_{a'}(Y[a'] \mid X)$ almost surely

Relaxation of assumptions

- Furthermore, I discuss the relaxation of **(3) Unconfoundedness** and **(2) Strong overlap** that yield **partial identification**
- These relaxations may invoke other assumptions in forms of **sensitivity models** (spectrum between point and partial identification)

Introduction: Spectrum of causal estimands

Layers of causation & heterogeneity

	Population level		Covariate level	
$\mathcal{L}_2$ **interventional**	average POs	$\mathbb{E}(Y[a])$	cond. average POs	$\mathbb{E}(Y[a] \mid x)$
	distributions of POs	$\mathbb{P}(Y[a])$	cond. distributions of POs	$\mathbb{P}(Y[a] \mid x)$
$\mathcal{L}_3$ **counterfactual**	average TE	$\mathbb{E}(Y[a] - Y[a'])$	cond. average TE	$\mathbb{E}(Y[a] - Y[a'] \mid x)$
	distribution of TE	$\mathbb{P}(Y[a] - Y[a'])$	cond. distribution of TE	$\mathbb{P}(Y[a] - Y[a'] \mid x)$
	average counterfactual outcome (ACO)	$\mathbb{E}(Y[a] \mid y'[a'])$		

- L2: What would the outcome be, on average or as a distribution, if the treatment were externally set to A ← a?
- L3: How would the units' outcomes have differed had they received a different treatment?
- Identification & estimation invoke different tools:
  - APOs/CAPOs/ATE/CATE → POF & the standard causal ML toolbox
  - DPOs/CDPOs → POF & probabilistic models toolbox
  - ACO/DTE/CDTE → POF + BGM / partial identification & custom probabilistic models

Introduction: Can we avoid causal assumptions?

Non-identifiability issue

- Main reason for causal assumptions: Data comes from $\mathcal{L}_1$, estimands live on $\mathcal{L}_2/\mathcal{L}_3$
  By the **causal hierarchy theorem** [1]: inference for a layer requires data from the same or a higher layer.
- W/o assumptions, we face the issue of **non-identifiability**:

$$\mathbb{P}^{\mathcal{M}^*}(X, A, Y) = \mathbb{P}^{\mathcal{M}'}(X, A, Y) \quad \text{and} \quad \mathbb{P}^{\mathcal{M}^*}(X, A, Y[a]) \neq \mathbb{P}^{\mathcal{M}'}(X, A, Y[a])$$

  the same $\mathcal{L}_1$ distributions — different $\mathcal{L}_2$ distributions

- Solutions:
  - Point identification → Adding assumptions: POF, BGMs, etc. But: the credibility of inference decreases
  - Partial identification → Inferring the whole interval, consistent with the data. But: might be non-informative (e.g., Manski bounds on the ATE/CATE always contain zero)

Law of decreasing credibility (Manski, 2003 [2])**:** The credibility of inference decreases with the strength of the assumptions maintained.

[1] Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In Probabilistic and Causal Inference: The Works of Judea Pearl, 507–556. ACM.

[2] Manski, C. F. (2003). Partial identification of probability distributions, volume 5. Springer.

Introduction: Three sources of uncertainty

Taxonomy of uncertainty sources

- Causal ML distinguishes three types of uncertainty:
  - **EU** (epistemic uncertainty): Finite-sample uncertainty of an estimator / learner
  - **AU** (aleatoric uncertainty): Intrinsic randomness of the causal target itself
  - **UU** (uncertainty due to unobserved covariates): Lack of identifiability due to missing variables
- Traditional ML usually distinguishes only EU and AU (UU appears rarely, e.g., as an omitted-variable bias)
- In causal ML:
  - EU is (relatively) well-studied
  - AU / UU / combinations are being actively studied

Introduction: Three sources of uncertainty

Taxonomy of uncertainty sources

- Causal ML distinguishes three types of uncertainty:
  - **EU** (epistemic uncertainty): Finite-sample uncertainty of an estimator / learner
  - **AU** (aleatoric uncertainty): Intrinsic randomness of the causal target itself
  - **UU** (uncertainty due to unobserved covariates): Lack of identifiability due to missing variables
- Traditional ML usually distinguishes only EU and AU (UU appears rarely, e.g., as an omitted-variable bias)
- In causal ML:
  - EU is (relatively) well-studied
  - AU / UU / combinations are being actively studied

How to reduce? (informally)

EU: Collect more data

AU: Collect more covariates

UU: Collect more covariates that provide identifiability

Introduction: Three sources of uncertainty

Taxonomy

- Causal ML distinguishes three types of uncertainty:
  - **EU** (epistemic uncertainty): Finite-sample uncertainty of an estimator / learner
  - **AU** (aleatoric uncertainty): Intrinsic randomne al target itself
  - **UU** (uncertainty due to unobserved covariates): Lack of identifiability due to missing variables

AU and UU are different!

How to reduce? (informally)

EU: Collect more data

AU: Collect more covariates

UU: Collect more covariates that provide identifiability

- Traditional ML usually distinguishes only EU and AU (UU appears rarely, e.g., as an omitted-variable bias)

- In causal ML:
  - EU is (relatively) well-studied
  - AU / UU / combinations are being actively studied

Introduction: Three sources of uncertainty

Definitions & examples

- **EU** (estimation / finite sample uncertainty):
  - Sampling variability of the estimator, $\hat{\theta}_n = \theta(\mathbb{P}_n)$, around the ground-truth $\theta = \theta(\mathbb{P})$
  - Example quantifiers: estimator variance $\mathrm{Var}(\hat{\theta}_n)$, estimator bias $\hat{\theta}_n - \theta$
  - Generally reducible with growing data size
- **AU** (target randomness):
  - Randomness of the target itself, remains even with the known best predictor (e.g., CAPOs/CATE)
  - Example quantifiers: variance of POs $\mathrm{Var}(Y[a] \mid x)$, CDFs of POs $F_a(y \mid x) = \mathbb{P}(Y[a] \leq y \mid x)$
  - Reducible by adding predictive covariates
- **UU** (lack of identifiability):
  - Observational data, $\mathbb{P}(Z)$, is compatible with several DGPs implying different θ
  - Example quantifiers: ignorance region $\theta \in [\theta^L, \theta^U]$
  - Reducible by adding assumptions (e.g., sensitivity models)
- All sources can interact (e.g., EU is always present while quantifying others)

Introduction: (Non-exhaustive) overview of approaches

Source	Estimand / target	Identification	Estimation
EU	asymptotic variance of A-IPTW; APO/ATE posterior	Point: POF	plug-in / robust estimators; prior & posterior corrections
AU	DPOs; CDPOs	Point: POF	plug-in; semi-parametric density estimation; GDR-learners
UU	confounded APOs/ATE, CAPOs/CATE; ACO	Partial: Manski bounds; sensitivity models (MSM, etc.)	plug-in; A-IPTW; TMLE; B-learner
AU + UU	CDF / quantiles / variance of the TE	Partial: Makarov, Fréchet–Hoeffding bounds	plug-in; AU-learners
EU + AU **EU + UU** **All three**	posterior predictive distribution / prediction intervals for POs and TEs	Point: POF; conformal prediction; Partial: MSM	amortized Bayesian inference; conformal prediction; Bayesian/conformal sensitivity analysis

Agenda

Introduction

Epistemic uncertainty:

- Plug-in estimators & Plug-in bias
- Debiasing with efficient influence functions
- Are the EIF-based estimators optimal?
- Finite-sample uncertainty
- Bayesian estimators
- Functional estimands

Aleatoric uncertainty

Uncertainty due to unobserved variables

Combining uncertainties

Takeaways

Epistemic uncertainty: Plug-in estimators & Plug-in bias

Plug-in estimators

- Under the assumptions (1)-(3) of POF, APOs/ATE are functionals of the observational distribution:

$$\theta_a(\mathbb{P}) = \mathbb{E}[\mu_a(X)], \theta_{a,a'}(\mathbb{P}) = \mathbb{E}[\mu_a(X) - \mu_{a'}(X)]$$

- We can consider the corresponding naïve plug-in estimators (that use nuisance function estimators):

$$\hat{\theta}^{\text{PI}}_{a,n} = \mathbb{P}_n\{\hat{\mu}_a(X)\}, \qquad \hat{\theta}^{\text{PI}}_{a,a',n} = \mathbb{P}_n\{\hat{\mu}_a(X) - \hat{\mu}_{a'}(X)\}$$

- However, the nuisance errors propagate in the **first-order**:

$$\hat{\theta}^{\text{PI}}_{a,n} - \theta_a = \underbrace{(\mathbb{P}_n - \mathbb{E})\{\mu_a(X)\}}_{\text{(I) CLT term}} + \underbrace{\mathbb{E}[\hat{\mu}_a(X) - \mu_a(X)]}_{\text{(II) first-order nuisance error}} + o_{\mathbb{P}}(1/\sqrt{n})$$

- Quantifying uncertainty (e.g., Wald confidence intervals given by CLT) requires
  - the knowledge of Term (II) → inaccessible
  - using such an estimator so that Term (II) converges faster-than-$\sqrt{n}$ → impossible

Epistemic uncertainty: Debiasing with efficient influence functions

- The EIF is the gradient of the functional along a smooth one-dimensional parametric submodel. E.g., Dirac-delta perturbation submodel, $\mathbb{P}_t(Z = z) = (1 - t)\mathbb{P}(Z = z) + t\delta(Z - z)$ reveals the EIF as:

$$\varphi_\theta(Z; \eta) := \frac{d}{dt}\left[\theta(\mathbb{P}_t)\right]\Big|_{t=0}$$

A-IPTW / TMLE estimators

- The EIF helps to debias plug-in estimators (i.e., one-step bias-correction). For example, the EIF for APOs is:

$$\varphi_a(Z; \eta) = \frac{1\{A = a\}}{\pi_a(X)}\left(Y - \mu_a(X)\right) + \mu_a(X) - \mathbb{E}[\mu_a(X)]$$

- Examples of debiased estimators are A-IPTW and TMLE estimators:

$$\hat{\theta}^{\text{A-IPTW}}_{a,n} = \mathbb{P}_n\left\{\frac{1\{A = a\}}{\hat{\pi}_a(X)}(Y - \hat{\mu}_a(X)) + \hat{\mu}_a(X)\right\}$$

$$\hat{\theta}^{\text{TMLE}}_{a,n} = \mathbb{P}_n\{\tilde{\mu}_a(X)\},\ \text{s.t. } \mathbb{P}_n\{\varphi_a(Z; \tilde{\eta})\} = 0$$

- Debiased estimators are **semi-parametric efficient** & (often) **doubly-robust**

Epistemic uncertainty: Debiasing with efficient influence functions

A-IPTW / TMLE estimators

- Now, the errors of the nuisance functions propagate with the **second-order**:

$$\hat{\theta}_n^{\text{A-IPTW}} - \theta = \underbrace{(\mathbb{P}_n - \mathbb{E})\{\varphi_\theta(Z;\eta)\}}_{\text{(I)}} + \underbrace{R_2(\hat{\eta},\eta)}_{\text{(II) second order}} + o_{\mathbb{P}}(1/\sqrt{n}), \quad R_2 = O_{\mathbb{P}}\big(\|\hat{\mu}_a - \mu_a\|_{L_2}\,\|1/\hat{\pi}_a - 1/\pi_a\|_{L_2}\big)$$

- Quantifying uncertainty (e.g., Wald confidence intervals given by CLT) requires:
  - using such an estimator so that Term (II) converges faster-than-$\sqrt{n}$ → when each nuisance is estimated with rate at least $o_{\mathbb{P}}(n^{-1/4})$
  - correct sample-splitting / Donsker conditions → possible
- CLT yields **asymptotic normality:**

$$\sqrt{n}\big(\hat{\theta}_n^{\text{A-IPTW}} - \theta\big) \rightsquigarrow \mathcal{N}(0,\sigma^2), \quad \sigma^2 = \mathrm{Var}[\varphi_\theta(Z;\eta)]$$

$$\mathrm{CI}_{1-\alpha}(\mathcal{D}) = \hat{\theta}_n^{\text{A-IPTW}} \pm z_{1-\alpha/2}\,\hat{\sigma}/\sqrt{n}$$

- The asymptotic variance is usually estimated with a plug-in estimator (not doubly-robust itself) or sandwich estimators (robust to propensity score misspecification) (Wu et al., 2025 [1])

[1] Wu, Hao, et al. "Why Is the Double-Robust Estimator for Causal Inference Not Doubly Robust for Variance Estimation?." arXiv preprint arXiv:2511.17907 (2025)..

Epistemic uncertainty: Are the EIF-based estimators optimal?

Local vs. global optimality

- **Semi-parametric efficiency** is a **local optimality criterion** (achieved by EIF-based estimators): best estimator in shrinking neighbourhoods of the ground truth distributions:

$$\inf_{\delta>0} \liminf_{n\to\infty} \sup_{d_{\mathrm{TV}}(\mathbb{P};\mathbb{Q})<\delta} n\,\mathbb{E}_{\mathcal{D}\sim\mathbb{Q}^n}\left[(\hat{\psi}_n - \psi(\mathbb{Q}))^2\right] \geq \mathrm{Var}[\varphi_\psi(Z;\eta)] = \sigma^2$$

  This inequality is known as a **semi-parametric efficiency bound**

- **Minimax optimality** is a **global optimality criterion:** best estimator wrt. worst-case risk over a global class of distributions:

$$\mathcal{R}_n(\mathcal{P}) = \inf_{\hat{\psi}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathcal{D}\sim\mathbb{P}^n}\left[(\hat{\psi}_n - \psi(\mathbb{P}))^2\right]$$

  In general, minimax optimal estimators **differ** from the semi-parametric efficient ones

- For APOs/ATE with black-box nuisances the **two views align**: EIF-based estimators are structure-agnostic minimax optimal up to constants

- **"Debias–variance" trade-off** (for general parameters): higher-order corrections shrink term (II) but potentially inflate term (I) and complexity → there is **no simple way** to construct minimax optimal estimators

Epistemic uncertainty: Finite-sample uncertainty

- EIF-based estimators only provide an asymptotic confidence intervals (via asymptotic normality). That is, **asymptotic validity** holds:

$$\liminf_n \mathbb{P}_{\mathcal{D}}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha$$

- Can we construct a non-trivial (finite-length) confidence interval with **finite-sample validity**?

$$\mathbb{P}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha \quad \text{for every n}$$

Finite-sample uncertainty

  - Only POF:

[1] Maclaren, O. J., & Nicholson, R. (2019). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv:1904.02826.

Epistemic uncertainty: Finite-sample uncertainty

- EIF-based estimators only provide an asymptotic confidence intervals (via asymptotic normality). That is, **asymptotic validity** holds:

$$\liminf_n \mathbb{P}_{\mathcal{D}}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1 - \alpha$$

- Can we construct a non-trivial (finite-length) confidence interval with **finite-sample validity**?

$$\mathbb{P}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1 - \alpha \quad \text{for every n}$$

  - Only POF: **No!** (we can make propensity score arbitrary complex for fixed n)

Finite-sample uncertainty

[1] Maclaren, O. J., & Nicholson, R. (2019). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv:1904.02826.

Epistemic uncertainty: Finite-sample uncertainty

- EIF-based estimators only provide an asymptotic confidence intervals (via asymptotic normality). That is, **asymptotic validity** holds:

$$\liminf_n \mathbb{P}_{\mathcal{D}}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1 - \alpha$$

- Can we construct a non-trivial (finite-length) confidence interval with **finite-sample validity**?

$$\mathbb{P}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1 - \alpha \quad \text{for every n}$$

Finite-sample uncertainty

  - Only POF: **No!** (we can make propensity score arbitrary complex for fixed n)
  - POF + known propensity score:

[1] Maclaren, O. J., & Nicholson, R. (2019). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv:1904.02826.

Epistemic uncertainty: Finite-sample uncertainty

- EIF-based estimators only provide an asymptotic confidence intervals (via asymptotic normality). That is, **asymptotic validity** holds:

$$\liminf_n \mathbb{P}_{\mathcal{D}}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha$$

- Can we construct a non-trivial (finite-length) confidence interval with **finite-sample validity**?

$$\mathbb{P}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha \quad \text{for every n}$$

Finite-sample uncertainty

  - Only POF: **No!** (we can make propensity score arbitrary complex for fixed n)
  - POF + known propensity score: **Still no!** (even sample mean doesn't have a finite-sample valid CI: we can infinitesimally perturb the observational distribution yet change the mean)

[1] Maclaren, O. J., & Nicholson, R. (2019). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv:1904.02826.

Epistemic uncertainty: Finite-sample uncertainty

- EIF-based estimators only provide an asymptotic confidence intervals (via asymptotic normality). That is, **asymptotic validity** holds:

$$\liminf_n \mathbb{P}_{\mathcal{D}}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha$$

- Can we construct a non-trivial (finite-length) confidence interval with **finite-sample validity**?

$$\mathbb{P}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha \quad \text{for every n}$$

Finite-sample uncertainty

  - Only POF: **No!** (we can make propensity score arbitrary complex for fixed n)
  - POF + known propensity score: **Still no!** (even sample mean doesn't have a finite-sample valid CI: we can infinitesimally perturb the observational distribution yet change the mean)
  - POF + known propensity score + bounded outcome:

[1] Maclaren, O. J., & Nicholson, R. (2019). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv:1904.02826.

Epistemic uncertainty: Finite-sample uncertainty

- EIF-based estimators only provide an asymptotic confidence intervals (via asymptotic normality). That is, **asymptotic validity** holds:

$$\liminf_n \mathbb{P}_{\mathcal{D}}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha$$

- Can we construct a non-trivial (finite-length) confidence interval with **finite-sample validity**?

$$\mathbb{P}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha \quad \text{for every n}$$

Finite-sample uncertainty

- Only POF: **No!** (we can make propensity score arbitrary complex for fixed n)
- POF + known propensity score: **Still no!** (even sample mean doesn't have a finite-sample valid CI: we can infinitesimally perturb the observational distribution yet change the mean)
- POF + known propensity score + bounded outcome: **Yes!** (Hoeffding bound)

[1] Maclaren, O. J., & Nicholson, R. (2019). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv:1904.02826.

Epistemic uncertainty: Finite-sample uncertainty

- EIF-based estimators only provide an asymptotic confidence intervals (via asymptotic normality). That is, **asymptotic validity** holds:

$$\liminf_n \mathbb{P}_{\mathcal{D}}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha$$

- Can we construct a non-trivial (finite-length) confidence interval with **finite-sample validity**?

$$\mathbb{P}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha \quad \text{for every n}$$

Finite-sample uncertainty

- Only POF: **No!** (we can make propensity score arbitrary complex for fixed n)
- POF + known propensity score: **Still no!** (even sample mean doesn't have a finite-sample valid CI: we can infinitesimally perturb the observational distribution yet change the mean)
- POF + known propensity score + bounded outcome: **Yes!** (Hoeffding bound)
- POF + known propensity score + sub-Gaussian tails for the outcome: **Yes!** (sub-Gaussian concentration inequality)
- POF + Hölder smooth nuisance functions + bounded outcome: **Yes!**

[1] Maclaren, O. J., & Nicholson, R. (2019). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv:1904.02826.

Epistemic uncertainty: Finite-sample uncertainty

- EIF-based estimators only provide an asymptotic confidence intervals (via asymptotic normality). That is, **asymptotic validity** holds:

$$\liminf_n \mathbb{P}_{\mathcal{D}}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha$$

- Can we construct a non-trivial (finite-length) confidence interval with **finite-sample validity**?

$$\mathbb{P}(\theta \in \mathrm{CI}_{1-\alpha}(\mathcal{D})) \geq 1-\alpha \quad \text{for every n}$$

Finite-sample uncertainty

  - Only POF: **No!** (we can make propensity score arbitrary complex for fixed n)
  - POF + known propensity score: **Still no!** (even sample mean doesn't have a finite-sample valid CI: we can infinitesimally perturb the observational distribution yet change the mean)
  - POF + known propensity score + bounded outcome: **Yes!** (Hoeffding bound)
  - POF + known propensity score + sub-Gaussian tails for the outcome: **Yes!** (sub-Gaussian concentration inequality)
  - POF + Hölder smooth nuisance functions + bounded outcome: **Yes!**

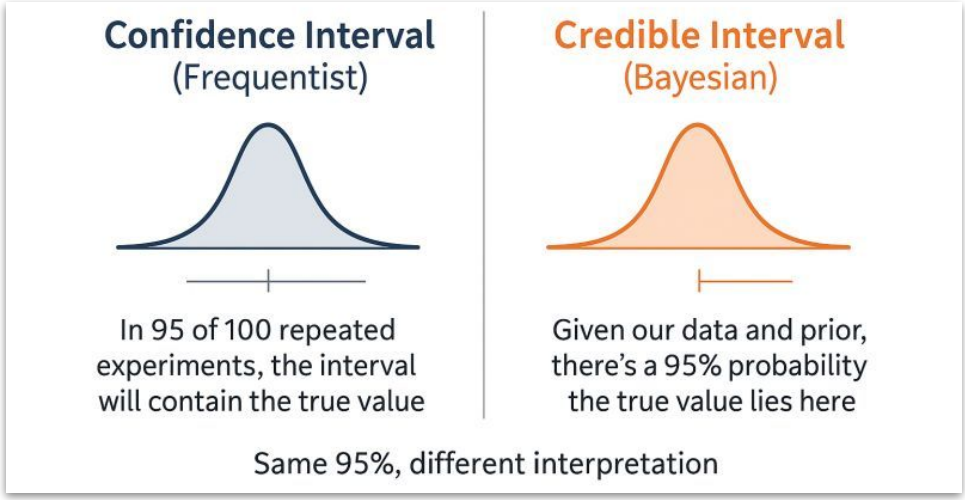
- The impossibility result highlights the difference between **identifiability** and **estimability:** Even with the established identifiability we **cannot** perform finite-sample inference (Maclaren et al., 2019 [1])

[1] Maclaren, O. J., & Nicholson, R. (2019). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv:1904.02826.

Epistemic uncertainty: Bayesian estimators

Bayesian estimators

- Alternatively, finite-sample uncertainty can be obtained “out-of-the-box” by employing Bayesian estimators
- Bayesian estimators proceed as follows (e.g., **plug-in posterior** for APOs):
  - Put priors on nuisance functions, then infer posteriors: $\tilde{\mu}_a \sim \Pi(\mu_a \mid \mathcal{D})$
  - Aggregate the posteriors with Bayesian bootstrap over $\mathbb{P}(X)$: $\theta_a(\tilde{\eta}) \mid \mathcal{D} = \mathrm{BB}_n\{\tilde{\mu}_a(X)\}$
  - Report credibility intervals, given the sample from posterior
- In general, credibility intervals are **not equal** to confidence intervals!
- In causal ML setting, even **asymptotically**. This happens due to **prior-induced confounding**: priors are too informative and keep an asymptotic influence on posterior
- To recover the asymptotic correspondence (= Bernstein–von Mises theorem), EIF-based **prior/posterior/VB corrections** were proposed (analogous to A-IPTW/TMLE estimators) (Ray et al., 2020; Yiu et al. 2025; Javurek et al., 2026 [1])



[1] Ray, K., & van der Vaart, A. (2020). Semiparametric Bayesian causal inference. The Annals of Statistics 48(5), 2999–3020. · Yiu, Andrew, et al. "Semiparametric posterior corrections." Journal of the Royal Statistical Society Series B: Statistical Methodology 87.4 (2025): 1025-1054. · Javurek, E., Frauen, D., Wang, Y., & Feuerriegel, S. (2026). Generalized Bayes for Causal Inference. arXiv preprint arXiv:2603.03035.

Epistemic uncertainty: Functional estimands

- For CAPOs/CATE the EIF is not well defined. Namely, it contains a Dirac delta, so debiased estimators do not exist for continuous covariates:

$$\varphi_{\theta_a(x)}(Z;\eta) = \frac{1\{A=a\}\,\delta(X-x)}{\pi_a(X)\,\mathbb{P}(X=x)}\,(Y-\mu_a(x))$$

- As a remedy, **orthogonal statistical learning** aims to find the best projection of CAPOs/CATE onto a target class $\mathcal{G} = \{g(v) : \mathcal{V} \to \mathcal{Y}\}$ by minimizing MSE risks, e.g.,

Meta-learners

$$\mathcal{L}_a(g,\hat{\eta}) = \mathbb{E}[(\hat{\mu}_a(X) - g(V))^2], \quad \mathcal{L}_{a,a'}(g,\hat{\eta}) = \mathbb{E}[(\hat{\mu}_a(X) - \hat{\mu}_{a'}(X) - g(V))^2]$$

- EIF-based estimators are now substituted by Neyman-orthogonal meta-learners: the gradient of the target risk is first-order insensitive to nuisance function errors. This yields familiar DR-/R-learners (quasi-oracle efficient)

- But orthogonality alone gives **no uncertainty bands** — those need bootstrap or specific target models with built-in EU quantification

Agenda

Introduction

Epistemic uncertainty

Aleatoric uncertainty:

- Why AU matters?
- AU quantifiers
- Distributional treatment effects

Uncertainty due to unobserved variables

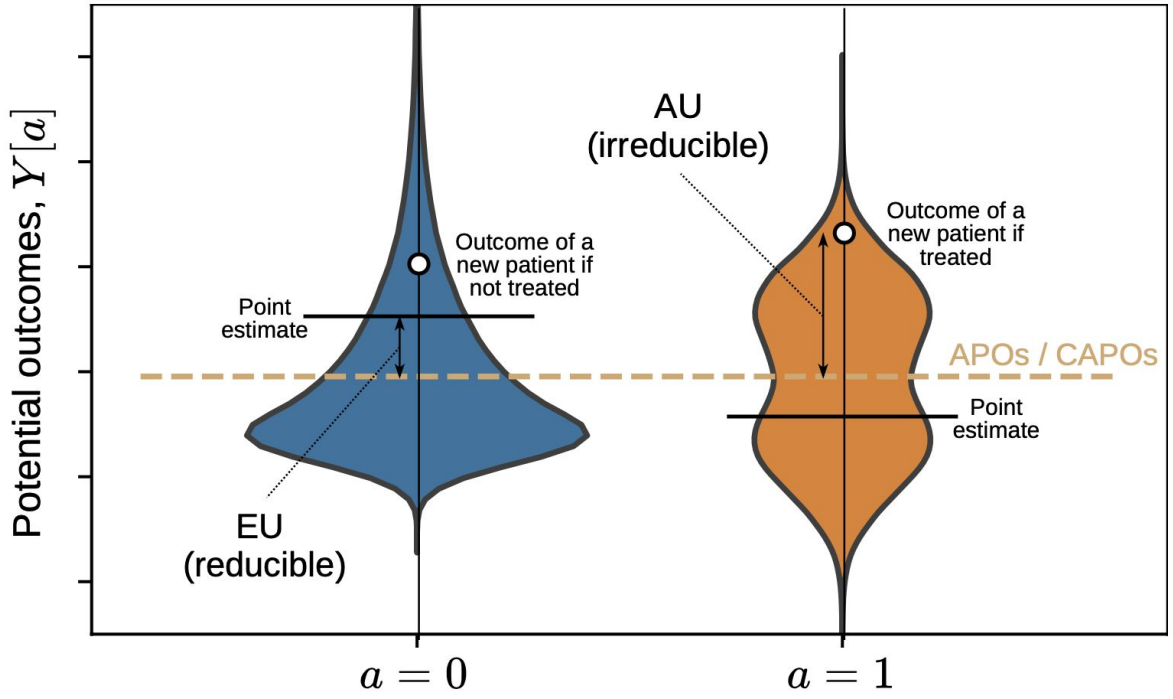
Combining uncertainties

Takeaways

Aleatoric uncertainty: Why AU matters?

Decision relevance

- AU is irreducible: It remains under correct identification and infinite data, because causal targets are heterogeneous beyond their means
- Two treatments may satisfy $\mathbb{E}(Y[a] \mid x) \approx \mathbb{E}(Y[a'] \mid x)$ while their CDPOs differ substantially
- One may have larger variance, heavier tails, or a higher probability of treatment failure



Aleatoric uncertainty: AU quantifiers

AU quantifiers

- Any functional of DPOs/CDPOs beyond the mean: CDFs/densities, quantiles, interval probabilities, variances, IQRs, tail risk (CVaR)
- In my thesis, I focused on densities of DPOs and CDPOs — they capture the full AU and scale to higher-dimensional outcomes:
  - **DPO densities**:
    - plug-in estimator (marginalization of the conditional outcome density)

$$\hat{\mathbb{P}}(Y[a] = y) = \mathbb{P}_n\{\hat{\mathbb{P}}(Y = y \mid a, X)\}$$

    - debiased projection-based estimators (in the spirit of meta-learners): finding the KL-projection on the probabilistic model class $\mathcal{G} = \{g(y) : \mathcal{Y} \to \mathbb{R}^+; \int_{\mathcal{Y}} g(y)\,\mathrm{d}y = 1\}$

$$\mathcal{L}_a(g, \eta) = -\mathbb{E}[\log g(Y[a])] = -\mathbb{E}\left[\int_{\mathcal{Y}} \log(g(y))\,\mathbb{P}(Y = y \mid a, X)\,\mathrm{d}y\right]$$

  - **CDPO densities:**
    - GDR-learners: general class of Neyman-orthogonal learners with any conditional generative model as a target (Melnychuk et al., 2026 [1])

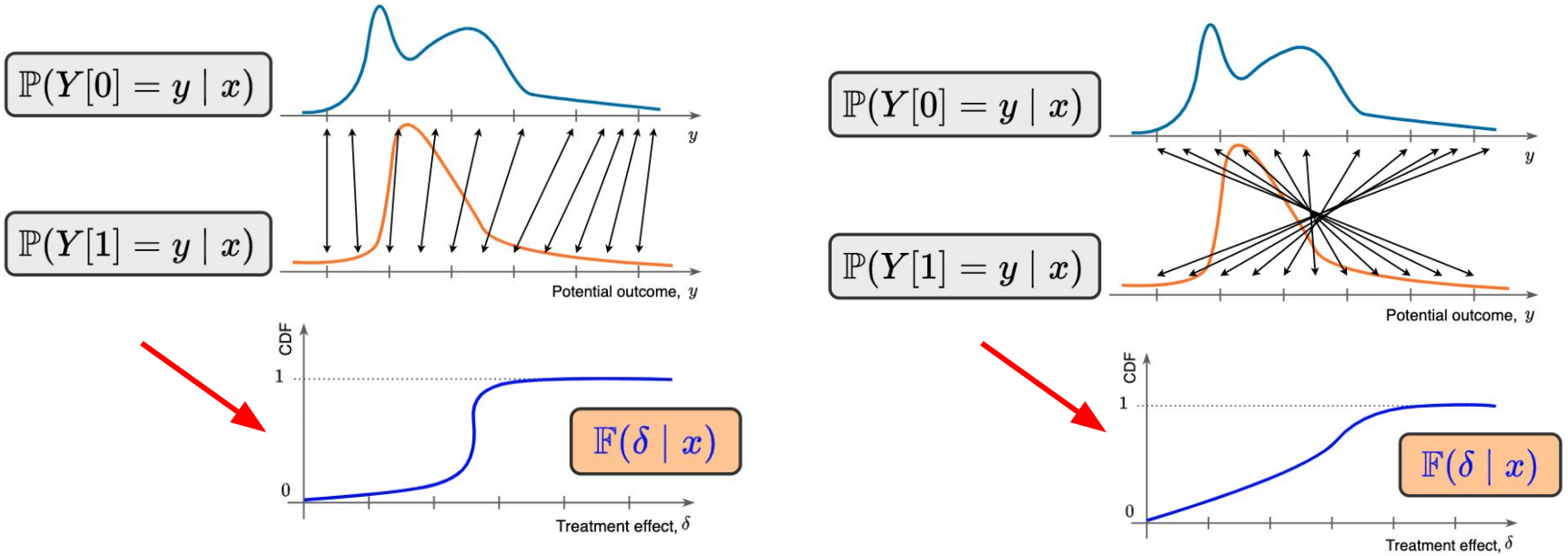
[1] Melnychuk, Valentyn, and Stefan Feuerriegel. "GDR-learners: Orthogonal learning of generative models for potential outcomes." International Conference on Learning Representations. Vol. 2026. 2026.

Aleatoric uncertainty: Distributional treatment effects

- We can go further and estimate the distributional contrasts: e.g., differences of quantiles $\mathbb{F}_{a}^{-1}(\alpha \mid x) - \mathbb{F}_{a'}^{-1}(\alpha \mid x)$. The distributional contrasts are also (confusingly) called **distributional treatment effects:** they compare marginal distributions of potential outcomes (thus, identifiable under POF)

- On the other hand, **distribution of the treatment effect** (DPO/CDPO), $\mathbb{P}(Y[a] - Y[a'] \mid x)$, quantifies the individual-level difference in the potential outcomes. DPO/CDPO depend on the joint distribution of potential outcomes (thus, unidentifiable under POF)

Dis-ambigua-tion



Agenda

Introduction

Epistemic uncertainty

Aleatoric uncertainty

Uncertainty due to unobserved variables:

- Hidden confounding
- Limited overlap
- Average counterfactual outcome

Combining uncertainties

Takeaways

Uncertainty due to unobserved variables: Hidden confounding

Relaxing (3) Unconfoundedness

- If we relax the (3) Unconfoundedness assumption, all the causal quantities are not point identified
- Recovering point identification (deconfounding, proximal causal inference) requires other strong, non-verifiable assumptions → rarely credible in practice
- (No-assumption) Manski bounds on CAPOs/APOs are valid, sharp but non-informative:

$$\psi_a^L(x) := \pi_a(x)\mu_a(x) + (1-\pi_a(x))\underline{y} \leq \mathbb{E}(Y[a] \mid x) \leq \psi_a^U(x) := \pi_a(x)\mu_a(x) + (1-\pi_a(x))\overline{y}$$

- Solution: a **sensitivity model** with a hyperparameter that spans from point identification to partial identification (no-assumption bounds). E.g., we can bound the ratios of propensities [1]:

$$\Gamma^{-1} \leq \frac{\pi_a(x) \,/\, (1-\pi_a(x))}{\pi_a^\star(x,u) \,/\, (1-\pi_a^\star(x,u))} \leq \Gamma \qquad \text{(marginal sensitivity model)}$$

  $\Gamma = 1$ recovers unconfoundedness; larger Γ allows for stronger hidden confounding

- Sharp CAPO/CATE bounds are adversarial tail averages (CVaRs) of the observed outcome law; APO/ATE bounds follow by marginalization
- *Estimation / meta-learning*: plug-in, A-IPTW estimators, orthogonal learners (B-learner)

[1] Dorn, J., Guo, K., & Kallus, N. (2025). Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. JASA 120(549), 331–342. · Oprescu, M. et al. (2023). B-learner: Quasi-oracle bounds on heterogeneous causal effects. ICML. · Frauen, D., Melnychuk, V., & Feuerriegel, S. (2023). Sharp bounds for generalized causal sensitivity analysis. NeurIPS . · Frauen, D., Imrie, F., Curth, A., Melnychuk, V., Feuerriegel, S., & van der Schaar, M. (2024). A neural framework for generalized causal sensitivity analysis. ICLR.

Uncertainty due to unobserved variables: Hidden confounding

Universe of sensitivity models

Sensitivity Model	Quantification of unmeasured confounding U	Examples
Effect Differences	$\lvert\psi_* - \psi\rvert$	Luedtke et al. [2015]
Potential Outcome Regression Differences	$\left\lVert \mathbb{E}(Y^a \mid A=a, X) - \mathbb{E}(Y^a \mid A=1-a, X) \right\rVert_p$	Díaz and van der Laan [2013] Luedtke et al. [2015]
Potential Outcome Regression Ratios	$\left\lVert \frac{\mathbb{E}(Y^a \mid A=1-a, X)}{\mathbb{E}(Y^a \mid A=a, X)} \right\rVert_p$	Lu and Ding [2023], Luedtke et al. [2015]
Direct Propensity Score Differences	$\lVert \pi_1(X) - \pi_1(X,W) \rVert_\infty$	Masten and Poirier [2018]
Propensity Score Odds Ratios	$\left\lVert \frac{\text{odds}\{\pi_1(X)\}}{\text{odds}\{\pi_1(X,W)\}} \right\rVert_p$ or $\left\lVert \frac{\text{odds}\{\pi_1(X,\widetilde{W})\}}{\text{odds}\{\pi_1(X,W)\}} \right\rVert_p$	Rosenbaum [2002], Tan [2006]
Explained Variance	$\frac{\lVert \mu_A(X,W) - \mu_A(X) \rVert_2^2}{\lVert Y - \mu_A(X) \rVert_2^2}$ and $\frac{\mathbb{E}\left(\left\{\pi_1(X,W)\pi_0(X,W)\right\}^{-1}\right)}{\mathbb{E}\left(\left\{\pi_1(X)\pi_0(X)\right\}^{-1}\right)}$	Chernozhukov et al. [2022] — assuming $\psi_* = \mathbb{E}(Y^1 - Y^0)$ (see Remark 3) Cinelli and Hazlett [2020]
Explained Variance	$\left\lVert \frac{\mathbb{V}[\text{odds}\{\pi_1(X,W))\}]}{\mathbb{V}[\text{odds}\{\pi_1(X)\}]} \right\rVert_\infty$	Huang and Pimentel [2022]

- The overview of existing sensitivity models for hidden confounding (borrowed from McClean et al. 2026 [1])

[1] McClean, Alec, Zach Branson, and Edward H. Kennedy. "Calibrated sensitivity models." Biometrika 113.2 (2026): asag001.

Uncertainty due to unobserved variables: Limited overlap

Relaxing (2) Strong overlap

- Alternatively, we can relax the (2) Strong overlap, by discarding the observations with too low / too high propensity scores
- This helps to reduce the EU, yet it effectively results in additional missing outcomes (hence, UU)
- **Non-overlap sensitivity model** was developed by Susmann et al. 2025 [1]: Under fixed propensity trimming threshold $c \in \left[0, \frac{1}{2}\right]$, APOs/ATE can be bounded:

$$\mathbb{E}\left(Y^1\right) = \mathbb{E}\left\{Y^1 \mathbb{I}\left(\pi > c\right)\right\} + \mathbb{E}\left(Y^1 \mid \pi \leq c\right) \mathbb{P}\left(\pi \leq c\right) \text{ and}$$
$$\mathbb{E}\left(Y^0\right) = \mathbb{E}\left\{Y^0 \mathbb{I}(\pi < 1 - c)\right\} + \mathbb{E}\left(Y^0 \mid \pi \geq 1 - c\right) \mathbb{P}(\pi \geq 1 - c)$$

- *Estimation / meta-learning*: Under the non-overlap sensitivity model, semi-parametric estimators were developed for the smoothed APOs/ATE bounds

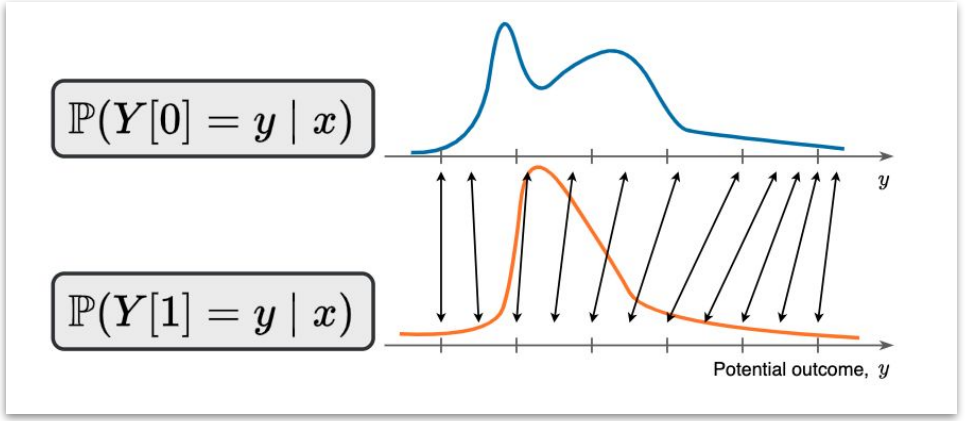
[1] Susmann, Herbert P., Alec McClean, and Iván Díaz. "Non-overlap Average Treatment Effect Bounds." arXiv preprint arXiv:2509.20206 (2025).

Uncertainty due to unobserved variables: Average counterfactual outcome

Average counter-factual outcome

- Average counterfactual outcomes (ACOs), $\mathbb{E}(Y[a] \mid y'[a'])$, are not identifiable under the POF
- To recover the point identification, we need BGM / rank invariance assumption:

$$\mathbb{F}_a(Y[a] \mid X) = \mathbb{F}_{a'}(Y[a'] \mid X) \text{ a.s.}$$



- We showed in Melnychuk et al., 2023 [1] that immediate relaxation of the BGM assumption (= latent noise is one-dimension higher than the outcome) yields **non-informative bounds** on ACOs
- We also suggested a remedy: a **curvature sensitivity model** (restricts the curvature of the level sets for functional assignment of the outcome wrt. to the latent noise)
- *Estimation / meta-learning*: Yet, neither sharp bounds were derived, nor there is an efficient estimator

[1] Melnychuk, Valentyn, Dennis Frauen, and Stefan Feuerriegel. "Partial counterfactual identification of continuous outcomes with a curvature sensitivity model." Advances in Neural Information Processing Systems 36 (2023): 32020-32060..

Agenda

Introduction

Epistemic uncertainty

Aleatoric uncertainty

Uncertainty due to unobserved variables

Combining uncertainties:

- AU + UU: Distribution of the treatment effect
- EU + AU: Posterior predictive distributions & Conformal prediction
- Other combinations

Takeaways

AU + UU: Distribution of the treatment effect

Partially identified AU

- Recap: The DTE/CDTE depend on the joint $\mathbb{P}(Y[a], Y[a'] \mid x)$, while the POF identify only the marginals $\mathbb{P}(Y[a] \mid x)$ $\rightarrow$ AU and UU appear together
- Partial identification results exist for different attributes of the DTE/CDTE:
  - CDF of DTE/CDTE is given by Makarov bounds:

$$\sup_{y \in \mathcal{Y}} \left[\mathbb{F}_a(y \mid x) - \mathbb{F}_{a'}(y - \delta \mid x)\right]_+ \leq \mathbb{F}_\Delta(\delta \mid x) \leq 1 + \inf_{y \in \mathcal{Y}} \left[\mathbb{F}_a(y \mid x) - \mathbb{F}_{a'}(y - \delta \mid x)\right]_-$$

  - variance of DTE/CDTE can be derived from Fréchet-Hoeffding bounds:

$$\int_0^1 F_1^{-1}(u) F_0^{-1}(1-u) du - \mu_0 \mu_1 \leq \mathrm{Cov}[Y(0), Y(1)] \leq \int_0^1 F_1^{-1}(u) F_0^{-1}(u) du - \mu_0 \mu_1$$

- *Estimation / meta-learning*: plug-in / IPTW / AU-learner (Melnychuk et al., 2024 [1]) for Makarov bounds; A-IPTW for Fréchet–Hoeffding

[1] Melnychuk, Valentyn, Stefan Feuerriegel, and Mihaela van der Schaar. "Quantifying aleatoric uncertainty of the treatment effect: a novel orthogonal learner." Advances in Neural Information Processing Systems 37 (2024): 105039-105089.

EU + AU: Posterior predictive distributions & Conformal prediction

AU equipped with EU

- The goal is to quantify both finite-sample uncertainty and the intrinsic randomness of potential outcomes
- Bayesian estimators: **posterior predictive distributions** (PPDs) of potential outcomes (Bayesian non-parametric models or amortized inference) (Balazadeh et al., 2025; Ma et al., 2025 [1]):

$$\mathbb{P}(Y[a] = y \mid x, \mathcal{D}) = \int \mathbb{P}_\theta(Y[a] = y \mid x)\, d\Pi(\theta \mid \mathcal{D}) = \int \mathbb{P}_\theta(Y = y \mid x, A = a)\, d\Pi(\theta \mid \mathcal{D}).$$

- Frequentist estimators: **conformal inference** that constructs **prediction intervals** for potential outcomes with finite-sample coverage guarantees (Lei & Candès, 2021 [2]):

$$\mathbb{P}\Big(Y[a] \in \widehat{C}_a(X)\Big) \geq 1 - \alpha$$

  - Note (Yang et al., 2024 [3]): Additional assumptions are needed! (The same finite-sample estimability issue reappears)

[1] Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C. Cress- well, and Rahul G. Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. In Advances in Neural Information Processing Systems, 2025. Ma, Yuchen, et al. "Foundation models for causal inference via prior-data fitted networks." International Conference on Learning Representations. Vol. 2026. 2026.

[2] Lihua Lei and Emmanuel J. Candès. Conformal Inference Of Counterfactuals And Individual treatment effects. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83 (5):911–938, 2021.

[3] Yang, Y., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. (2024). Doubly robust calibration of prediction sets under covariate shift. Journal of the Royal Statistical Society Series B: Statistical Methodology, 86(4), 943–965.

Other combinations

EU + UU

- Confidence intervals over partially identified estimands (e.g.. asymptotically normal CIs come “in one package” with A-IPTW estimators of MSM bounds (Dorn et al., 2025 [1]))
- Bayesian sensitivity analysis (prior on the nuisance functions / sensitivity parameter) (McCandless et al., 2007; Gustafson, 2010; Gustafson et al., 2010; Dhawan et al., 2026 [2])

EU + UU + AU

- Posterior predictive distribution of the treatment effect (Ma et al., 2026 [3]):

$$\mathbb{P}(Y[a] - Y[a'] \mid x, \mathcal{D})$$

- Conformal prediction of the treatment effect (Lei & Candès, 2021 [4])
- Conformal sensitivity analysis for POs/TE under the MSM (Yin et al., 2024 [5])

[1] Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. Journal of the American Statistical Association, 120(549):331–342, 2025.

[2] Lawrence C. McCandless, Paul Gustafson, and Adrian Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. Statistics in medicine, 26(11):2331–2347, 2007. · Gustafson, P. (2010). Bayesian inference for partially identified models. · Dhawan, Nikita, et al. "Bayesian Sensitivity of Causal Inference Estimators under Evidence-Based Priors." arXiv preprint arXiv:2605.07993 (2026)..

[3] Ma, Yuchen, et al. "Foundation models for causal inference via prior-data fitted networks." International Conference on Learning Representations. Vol. 2026. 2026.

[4] Lihua Lei and Emmanuel J. Candès. Conformal Inference Of Counterfactuals And Individual treatment effects. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83 (5):911–938, 2021.

[4] Mingzhang Yin, Claudia Shi, Yixin Wang, and David M. Blei. Conformal sensitivity analysis for individual treatment effects. Journal of the American Statistical Association, 119(545):122–135, 2024.

Agenda

Introduction

Epistemic uncertainty

Aleatoric uncertainty

Uncertainty due to unobserved variables

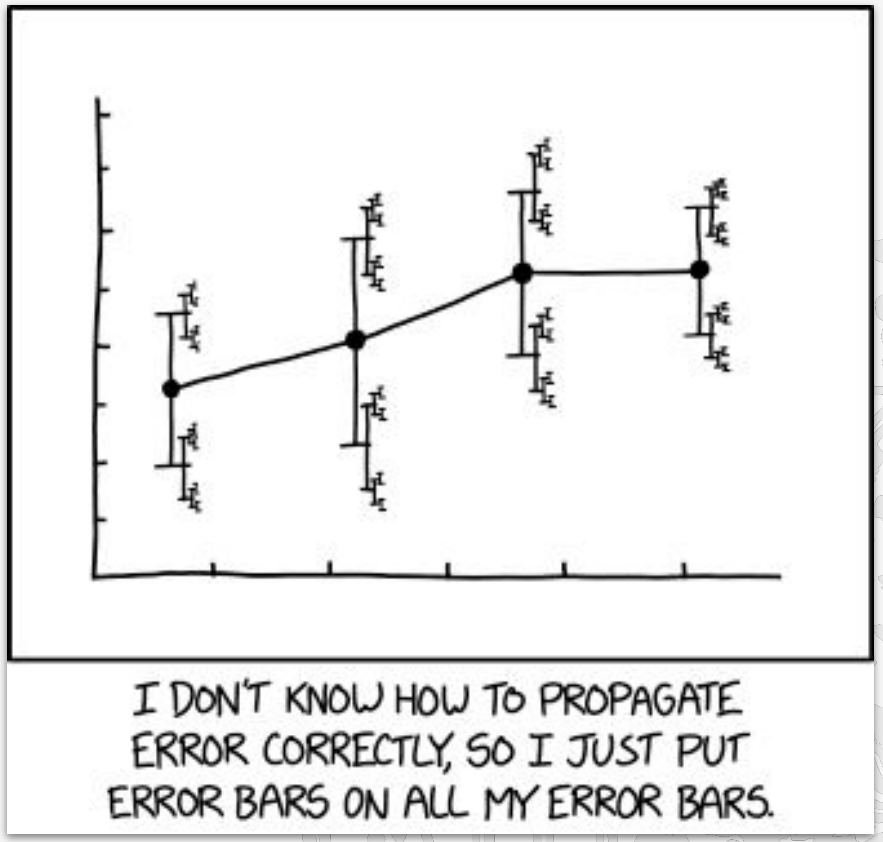
Combining uncertainties:

Takeaways

Takeaways

Uncertainty in causal ML is **not a single object**. Ask first *which* uncertainty the decision depends on — finite data (EU), intrinsic randomness (AU), or limited identifiability (UU) — and only then choose estimand, assumptions and estimator.

Thank you for attention!



https://www.explainxkcd.com/wiki/index.php/2110:_Error_Bars