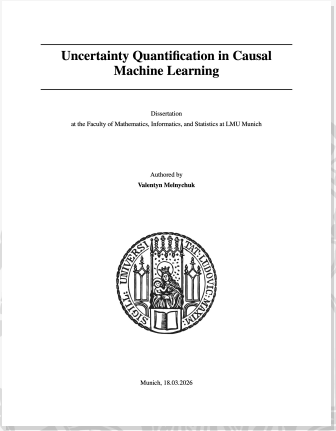
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN | LMU MUNICH SCHOOL OF MANAGEMENT | INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

# Uncertainty Quantification in Causal Machine Learning

**Valentyn Melnychuk**
LMU Munich | Faculty of Mathematics, Informatics and Statistics

PhD in Computer Science | Disputation
26.10.2026

Supervisor: Prof. Dr. Stefan Feuerriegel
Reviewers: Prof. Dr. Stefan Feuerriegel, Prof. Dr. Vasilis Syrgkanis, Prof. Dr. Rahul G. Krishnan
Organizing committee: Prof. Dr. Eyke Hüllermeier, Prof. Dr. Marcus Paradies



**Uncertainty Quantification in Causal Machine Learning**

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

Authored by
**Valentyn Melnychuk**

Munich, 18.03.2026

# Does “a glass of wine a day keep the doctor away”?



a glass of wine a day

?



health

Images: https://www.tiktok.com/@annaolala/video/7193719311150697733; Twemoji by Twitter, Inc. and contributors (CC BY 4.0)

# Does “a glass of wine a day keep the doctor away”?

**Observational data** ($\approx$ 30 cohort studies from 1981-2005, $>$1M people):

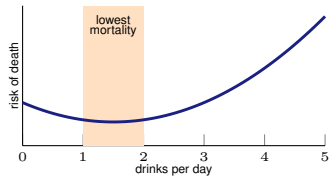
- number of alcoholic drinks per day (treatment $A$)
- death after 10 years (outcome $Y$)

**Aim:** effect of alcohol consumption on the risk of death

**Evidence:** light drinkers die *less often* than non-drinkers, namely, a J-shaped curve (Di Castelnuovo et al., 2006)

Conclusion: “1–2 drinks/day is the healthiest choice”

**J-shaped curve (schematic)**



# Does “a glass of wine a day keep the doctor away”?

**Observational data** ($\approx$ 30 cohort studies from 1981-2005, $>$1M people):

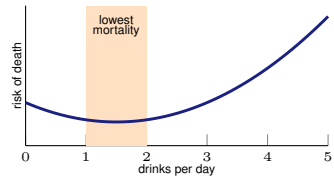
- number of alcoholic drinks per day (treatment $A$)
- death after 10 years (outcome $Y$)

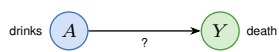
**Aim:** effect of alcohol consumption on the risk of death

**Evidence:** light drinkers die *less often* than non-drinkers, namely, a J-shaped curve (Di Castelnuovo et al., 2006)

Conclusion: “1–2 drinks/day is the healthiest choice”?

**J-shaped curve (schematic)**



# Does “a glass of wine a day keep the doctor away”?

**Observational data** ($\approx$ 30 cohort studies from 1981-2005, $>$1M people):
- number of alcoholic drinks per day (treatment $A$)
- death after 10 years (outcome $Y$)

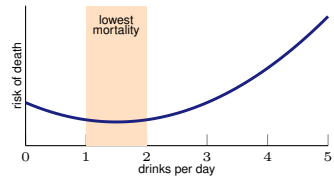
**Aim:** effect of alcohol consumption on the risk of death

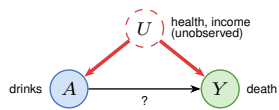
**Evidence:** light drinkers die *less often* than non-drinkers, namely, a J-shaped curve (Di Castelnuovo et al., 2006)

Conclusion: “1–2 drinks/day is the healthiest choice”?

**Problems with the study:**
- **unobserved confounding**: income; quitting *because of* poor health

**J-shaped curve (schematic)**



# Does “a glass of wine a day keep the doctor away”?

**Observational data** ($\approx$ 30 cohort studies from 1981-2005, $>$1M people):
- number of alcoholic drinks per day (treatment $A$)
- death after 10 years (outcome $Y$)

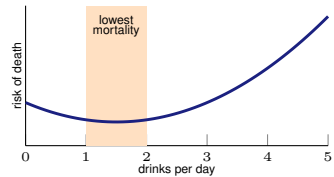
**Aim:** effect of alcohol consumption on the risk of death

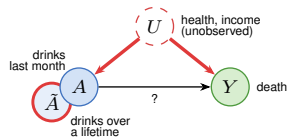
**Evidence:** light drinkers die *less often* than non-drinkers, namely, a J-shaped curve (Di Castelnuovo et al., 2006)

Conclusion: “1–2 drinks/day is the healthiest choice”?

**Problems with the study:**
- **unobserved confounding**: income; quitting *because of* poor health
- **violation of consistency**: e.g., “how much did you drink last month?” is not the same as “how much did you drink over a lifetime?”

**J-shaped curve (schematic)**



# Same question, different study designs, different answers

**Experimental data?**

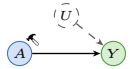
- randomize drinks per day
- follow up for 10 years

# Same question, different study designs, different answers

**Experimental data?**

- randomize drinks per day
- follow up for 10 years

**Attempted:** NIH trial, $>$ 7,800 people, one drink/day vs. none, cost: $100M



# Same question, different study designs, different answers

**Experimental data?**

- randomize drinks per day
- follow up for 10 years

**Attempted:** NIH trial, $>$ 7,800 people, one drink/day vs. none, cost: \$100M ($\approx$ \$67M of which from **alcohol companies**) $\Rightarrow$ compromised and terminated in 2018 (The New York Times, 2018)

# Same question, different study designs, different answers

**Experimental data?**

- randomize drinks per day
- follow up for 10 years

**Attempted:** NIH trial, $>$ 7,800 people, one drink/day vs. none, cost: \$100M ($\approx$ \$67M of which from **alcohol companies**) $\Rightarrow$ compromised and terminated in 2018 (The New York Times, 2018)

---

**Problem:** long-term drinking RCTs are costly (and, generally, unethical)

# Same question, different study designs, different answers

**Experimental data?**

- randomize drinks per day
- follow up for 10 years

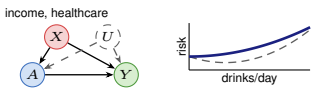
**Attempted:** NIH trial, $> 7{,}800$ people, one drink/day vs. none, cost: \$100M ($\approx$ \$67M of which from **alcohol companies**) $\Rightarrow$ compromised and terminated in 2018 (The New York Times, 2018)

**Problem:** long-term drinking RCTs are costly (and, generally, unethical)

**Observational data + covariates**

- adjust for socio-economic covariates $X$ (income, healthcare)
- separate former from lifetime non-drinkers, $A$ vs. $\tilde{A}$

**Evidence:** light drinkers have *no* lower mortality (Zhao et al., 2023)




**Problem:** it is impossible to fully rule out unobserved confounding

# Same question, different study designs, different answers

**Experimental data?**

- randomize drinks per day
- follow up for 10 years

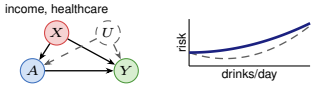
**Attempted:** NIH trial, $>$ 7,800 people, one drink/day vs. none, cost: \$100M ($\approx$ \$67M of which from **alcohol companies**) $\Rightarrow$ compromised and terminated in 2018 (The New York Times, 2018)

**Problem:** long-term drinking RCTs are costly (and, generally, unethical)

**Observational data + covariates**

- adjust for socio-economic covariates $X$ (income, healthcare)
- separate former from lifetime non-drinkers, $A$ vs. $\tilde{A}$

**Evidence:** light drinkers have *no* lower mortality (Zhao et al., 2023)

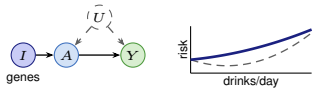



**Problem:** it is impossible to fully rule out unobserved confounding

**Observational data + instrument**

- genes $I$ that affect alcohol intake but not mortality directly (Mendelian randomization)

**Evidence:** *no* protective effect; cardiovascular risk rises with every level (Biddinger et al., 2022)




**Problem:** strong assumptions about the instrument

# Same question, different study designs, different answers

**Experimental data?**

- randomize drinks per day
- follow up for 10 years

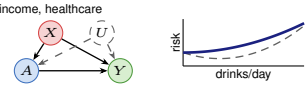
**Attempted:** NIH trial, $>7{,}800$ people, one drink/day vs. none, cost: \$100M ($\approx$ \$67M of which from **alcohol companies**) $\Rightarrow$ compromised and terminated in 2018 (The New York Times, 2018)

**Problem:** long-term drinking RCTs are costly (and, generally, unethical)

**Observational data + covariates**

- adjust for socio-economic covariates $X$ (income, healthcare)
- separate former from lifetime non-drinkers, $A$ vs. $\tilde{A}$

**Evidence:** light drinkers have *no* lower mortality (Zhao et al., 2023)

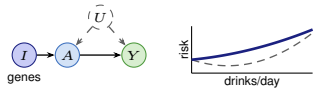



**Problem:** it is impossible to fully rule out unobserved confounding

**Observational data + instrument**

- genes $I$ that affect alcohol intake but not mortality directly (Mendelian randomization)

**Evidence:** *no* protective effect; cardiovascular risk rises with every level (Biddinger et al., 2022)




**Problem:** strong assumptions about the instrument

From causal evidence to policy: since January 2023, World Health Organization warns that **no level of alcohol consumption is safe for our health**

# Why do we need causal ML?

$\Rightarrow$ to move from **prediction** (*who is at risk?*) to **prescription** (*what works, and for whom?*)

### Traditional ML: prediction



**Umbrella problem:** will it rain? $\Rightarrow$ taking an umbrella
Accurate forecast $\Rightarrow$ good decision
Limitations: assuming that an action has no effect on the outcome

- Which patients are at high risk of cardiovascular events?
- Who is at risk of long-term unemployment?
- Who is likely to vote?

# Why do we need causal ML?

⇒ to move from **prediction** (*who is at risk?*) to **prescription** (*what works, and for whom?*)

### Traditional ML: prediction



**Umbrella problem:** will it rain? ⇒ taking an umbrella
Accurate forecast ⇒ good decision
Limitations: assuming that an action has no effect on the outcome

- Which patients are at high risk of cardiovascular events?
- Who is at risk of long-term unemployment?
- Who is likely to vote?

### Causal ML: prescription



**Rain dance problem:** dancing ⇒ will it rain?
Accurate knowledge of causal effect ⇒ good decision
Limitations: none

- Which patients benefit from an intensive treatment?
- Which training programme helps which job seeker?
- Whom would a canvassing visit mobilize?

# Why do we need causal ML?

$\Rightarrow$ to move from **prediction** (*who is at risk?*) to **prescription** (*what works, and for whom?*)

### Traditional ML: prediction



**Umbrella problem:** will it rain? $\Rightarrow$ taking an umbrella
Accurate forecast $\Rightarrow$ good decision
Limitations: assuming that an action has no effect on the outcome

- Which patients are at high risk of cardiovascular events?
- Who is at risk of long-term unemployment?
- Who is likely to vote?

### Causal ML: prescription



**Rain dance problem:** dancing $\Rightarrow$ will it rain?
Accurate knowledge of causal effect $\Rightarrow$ good decision
Limitations: none

- Which patients benefit from an intensive treatment?
- Which training programme helps which job seeker?
- Whom would a canvassing visit mobilize?

Accurate prediction **not always leads** to good decision. **Causal ML** – the overarching topic of my PhD – acknowledges that and aims to infer *what would happen to the outcome* $Y$ *if we set the treatment* $A \leftarrow a$.

# My PhD at a glance: Publications

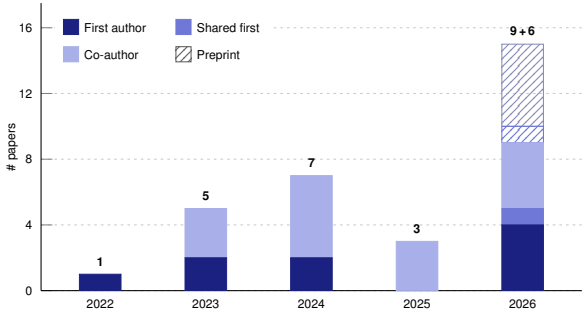
**25 + 6**
peer-reviewed publications
+ preprints (2021 – 2026)

**9 + 2**
first-author
+ shared-first-author papers

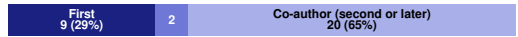
**20**
co-author papers
(second author or later)

**> 1,400**
citations
h-index 15

## Papers per year by author position




## Share by author position (31 papers incl. preprints)




11 of 31 papers (35%) first or shared-first author; peer-reviewed only: 10 of 25 (40%)

## Venues

ICLR 9 · NeurIPS 7 · ICML 4 · AISTATS 1 · AAAI 1
CLeaR 1 · Nature Medicine 1 · Wiley StatsRef 1 · Preprints 6

## Main topics

- uncertainty quantification in causal ML • causal sensitivity models
- time-varying treatment effects estimation • off-policy learning
- causal fairness • causal foundation models

# My PhD at a glance: First-author papers & beyond

| Year | Venue | Paper (first author) |
|---|---|---|
| 2022 | ICML | Causal Transformer for Estimating Counterfactual Outcomes |
| 2023 | ICML | Normalizing Flows for Interventional Density Estimation **Part I** |
| 2023 | NeurIPS† | Partial Counterfactual Identification of Continuous Outcomes with a Curvature Sensitivity Model **Part II** |
| 2024 | ICLR† | Bounds on Representation-Induced Confounding Bias for Treatment Effect Estimation **Part II** |
| 2024 | NeurIPS | Quantifying Aleatoric Uncertainty of the Treatment Effect: A Novel Orthogonal Learner **Part III** |
| 2026 | AISTATS‡ | Orthogonal Representation Learning for Estimating Causal Quantities |
| 2026 | ICLR | GDR-learners: Orthogonal Learning of Generative Models for Potential Outcomes **Part I** |
| 2026 | ICLR | Overlap-Adaptive Regularization for CATE Estimation |
| 2026 | ICML | Frequentist Consistency of Prior-Data Fitted Networks for Causal Inference |
| 2026 | CLeaR | Consistent End-to-End Estimation for Counterfactual Fairness (shared first) |
| 2026 | Under review | Debiased Prior-Data Fitted Networks for Amortized Bayesian Semi-Parametric Inference (shared first) |

† spotlight ‡ oral presentation **Part ⋆** = included in the dissertation

## Beyond papers

- **Research visits:** University of Cambridge (2024); University of Toronto & Vector Institute (2025)
- **Director** of the Causal ML Lab at LMU Munich (since 2024)
- Associated PhD student, **Konrad Zuse School relAI** (since 2023)
- **Reviewer awards:** AISTATS 2023, NeurIPS 2023 (top reviewer); ICML 2026 (gold reviewer); NeurIPS 2026 (top reviewer)
- **Tutorials & invited talks**, e.g., Munich Causal ML Workshops, Human Technopole, Vector Institute, University of Mannheim

# Dissertation: Methodological advances in three parts

## Part I: AU

Aleatoric Uncertainty



**Interventional normalizing flows (INFs)** (ICML 2023): density estimation for marginal distributions of potential outcomes

**Generative doubly-robust (GDR) learners** (ICLR 2026): orthogonal learning for conditional distributions of potential outcomes

today

## Part II: UU

Uncertainty due to Unobserved Variables



**Representation-induced confounding bias (RICB)** (ICLR 2024): unveiling and addressing the confounding in learned representations of covariates

**Curvature sensitivity model (CSM)** (NeurIPS 2023): introducing a sensitivity model for counterfactual inference of continuous outcomes

backup

## Part III: AU + UU

Combining both Uncertainties



**Aleatoric uncertainty (AU) learner** (NeurIPS 2024): orthogonal learning for bounds on the distribution of the treatment effect

today

# Dissertation: Methodological advances in three parts

## Part I: AU

Aleatoric Uncertainty



**Interventional normalizing flows (INFs)** (ICML 2023): density estimation for marginal distributions of potential outcomes

**Generative doubly-robust (GDR) learners** (ICLR 2026): orthogonal learning for conditional distributions of potential outcomes

today

## Part II: UU

Uncertainty due to Unobserved Variables



**Representation-induced confounding bias (RICB)** (ICLR 2024): unveiling and addressing the confounding in learned representations of covariates

**Curvature sensitivity model (CSM)** (NeurIPS 2023): introducing a sensitivity model for counterfactual inference of continuous outcomes

backup

## Part III: AU + UU

Combining both Uncertainties



**Aleatoric uncertainty (AU) learner** (NeurIPS 2024): orthogonal learning for bounds on the distribution of the treatment effect

today

# Agenda

**Introduction**

- Causal ML & uncertainty
- Problem formulation; Spectrum of causal estimands; Identification vs. estimation
- Debiased estimation
- Three sources of uncertainty; Case study (sources of uncertainty)

**Part I: AU (Aleatoric Uncertainty)**

**Part III: AU + UU (Combining Both Uncertainties)**

**Summary & outlook**

"Part II: UU (Uncertainty due to Unobserved Variables)" is covered in the backup slides

# Introduction: Causal ML & uncertainty

**What is causal ML?**

- **Causal ML** bridges structural causal reasoning (Rubin, 1974; Pearl, 2009) and flexible predictive ML models (Kuhn & Johnson, 2013)
- Prediction targets are inherently causal:
  - potential outcomes (POs): $Y[a]$
  - treatment effects (TEs): $Y[1] - Y[0]$
- Applications: personalized medicine, public policy, marketing, education; whenever randomized controlled trials (RCTs) are limited or infeasible

# Introduction: Causal ML & uncertainty

**What is causal ML?**

- **Causal ML** bridges structural causal reasoning (Rubin, 1974; Pearl, 2009) and flexible predictive ML models (Kuhn & Johnson, 2013)
- Prediction targets are inherently causal:
  - potential outcomes (POs): $Y[a]$
  - treatment effects (TEs): $Y[1] - Y[0]$
- Applications: personalized medicine, public policy, marketing, education; whenever randomized controlled trials (RCTs) are limited or infeasible

**Why uncertainty?**

- In sensitive domains (e.g., medicine), a point estimate is not enough $\rightarrow$ reliable deployment requires **quantifying the uncertainty** around it
- With different types of uncertainty, we can represent **limited knowledge** of the ground-truth causal target
- This increases **trustworthiness** and **reliability** of decision-making

# Introduction: Problem formulation

**Problem formulation**

Given i.i.d. observational data

$$\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}(X, A, Y):$$

- $X$ covariates (e.g., patient history)
- $A$ binary treatment, $A \in \{0, 1\}$
- $Y$ continuous outcome, $Y \in \mathbb{R}$

Goal: predict $Y[a]$ and $Y[1] - Y[0]$

| Patient | $X$ | $A$ | $Y[0]$ | $Y[1]$ | $Y[1] - Y[0]$ |
|---|---|---|---|---|---|
| 1 | $x_1$ | 0 | $-1.0$ | ? | ? |
| 2 | $x_2$ | 1 | ? | 2.3 | ? |
| 3 | $x_3$ | 1 | ? | 0.3 | ? |

We **never** observe both POs $\Rightarrow$ *missing data problem*

# Introduction: Problem formulation

**Problem formulation**

Given i.i.d. observational data $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n \sim \mathbb{P}(X, A, Y)$:

- $X$ covariates (e.g., patient history)
- $A$ binary treatment, $A \in \{0, 1\}$
- $Y$ continuous outcome, $Y \in \mathbb{R}$

Goal: predict $Y[a]$ and $Y[1] - Y[0]$

| Patient | $X$ | $A$ | $Y[0]$ | $Y[1]$ | $Y[1] - Y[0]$ |
|---|---|---|---|---|---|
| 1 | $x_1$ | 0 | $-1.0$ | ? | ? |
| 2 | $x_2$ | 1 | ? | 2.3 | ? |
| 3 | $x_3$ | 1 | ? | 0.3 | ? |

We **never** observe both POs $\Rightarrow$ *missing data problem*

---

**Nuisance functions**

To impute the missing POs, I later use the following functions $\eta$ (derived from the observational distribution $\mathbb{P}(X, A, Y)$):

- propensity score: $\pi_a(x) = \mathbb{P}(A = a \mid X = x)$
- conditional outcome expectation: $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$
- conditional outcome density / CDF: $\xi_a(y \mid x) = \mathbb{P}(Y = y \mid X = x, A = a)$, $\mathbb{F}_a(y \mid x) = \mathbb{P}(Y \leq y \mid X = x, A = a)$

These are called **nuisance functions**, as they are **not** the main targets of inference/learning (i.e., not estimands) but rather an **"obstacle"** needed for causal inference/causal ML.

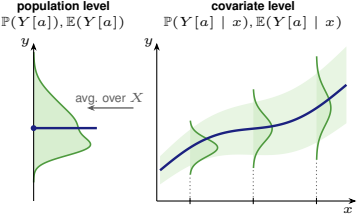
# Introduction: Spectrum of causal estimands

| Layer of causation | Population level | | Covariate level | |
|---|---|---|---|---|
| $\mathcal{L}_1$ **observational** | average outcomes | $\mathbb{E}(Y \mid A = a)$ | cond. average outcomes | $\mu_a(x) = \mathbb{E}(Y \mid x, A = a)$ |
| | distributions of outcomes | $\mathbb{P}(Y \mid A = a)$ | cond. outcome distributions | $\mathbb{P}(Y \mid x, A = a)$ |
| $\mathcal{L}_2$ **interventional** | average POs (APOs) | $\mathbb{E}(Y[a])$ | cond. average POs (CAPOs) | $\mathbb{E}(Y[a] \mid x)$ |
| | distributions of POs (DPOs) | $\mathbb{P}(Y[a])$ | cond. distributions of POs (CDPOs) | $\mathbb{P}(Y[a] \mid x)$ |
| $\mathcal{L}_3$ **counterfactual** | average TE (ATE) | $\mathbb{E}(Y[1] - Y[0])$ | cond. average TE (CATE) | $\mathbb{E}(Y[1] - Y[0] \mid x)$ |
| | distribution of TE (DTE) | $\mathbb{P}(Y[1] - Y[0])$ | cond. distribution of TE (CDTE) | $\mathbb{P}(Y[1] - Y[0] \mid x)$ |
| | avg. counterfactual outcome (ACO) | $\mathbb{E}(Y[a] \mid y'[a'])$ | | |

= causal estimands studied in this thesis (Part II, additionally studies representation-level CATE, $\mathbb{E}(Y[1] - Y[0] \mid \Phi(X) = \phi)$)

Causal ML aims to infer/learn $\mathcal{L}_2$ and $\mathcal{L}_3$ causal estimands from $\mathcal{L}_1$ data:

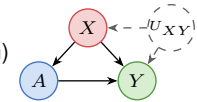
- $\mathcal{L}_2$ **interventional (potential)**: What would the outcome $Y$ be, on average or as a distribution, if the treatment were *set* to $A \leftarrow a$?
- $\mathcal{L}_3$ **counterfactual**: How would the *same units'* outcomes $y'$ have differed had they received a different treatment ($a$ vs. $a'$)?



# Introduction: Identification vs. estimation

**Identification**

- Can a causal estimand $\psi$ be written as a functional of $\mathbb{P}(X, A, Y)$ *(infinite observational data)*: $\psi = \psi(\mathbb{P})$? $\Rightarrow$ in general, no! $\Rightarrow$ we need additional **identification assumptions**
- I rely on the assumptions of a **potential outcomes framework (POF)** (Rubin, 1974):
  1. **Consistency:** $Y = Y[A]$ (= POs are partially observable)
  2. **Strong overlap:** $\varepsilon \leq \pi_a(X) \leq 1 - \varepsilon$ (= treatment is random enough)
  3. **Unconfoundedness:** $Y[a] \perp\!\!\!\perp A \mid X$ (= all confounders are in $X$)



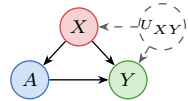

- Under Assumptions 1-3, all $\mathcal{L}_2$ causal estimands (and some $\mathcal{L}_3$) are identifiable:
  - CAPOs: $\mathbb{E}(Y[a] \mid x) = \mu_a(x)$; CATE: $\mathbb{E}(Y[1] - Y[0] \mid x) = \mu_1(x) - \mu_0(x)$
  - APOs: $\mathbb{E}(Y[a]) = \mathbb{E}[\mu_a(X)]$ (regression adjustment), $\cdots = \mathbb{E}[Y \mathbb{1}\{A = a\}/\pi_a(X)]$ (inverse probability of treatment weighting, IPTW)

# Introduction: Identification vs. estimation

**Identification**

- Can a causal estimand $\psi$ be written as a functional of $\mathbb{P}(X, A, Y)$ *(infinite observational data)*: $\psi = \psi(\mathbb{P})$? $\Rightarrow$ in general, no! $\Rightarrow$ we need additional **identification assumptions**
- I rely on the assumptions of a **potential outcomes framework (POF)** (Rubin, 1974):

  **1. Consistency:** $Y = Y[A]$ (= POs are partially observable)
  **2. Strong overlap:** $\varepsilon \leq \pi_a(X) \leq 1 - \varepsilon$ (= treatment is random enough)
  **3. Unconfoundedness:** $Y[a] \perp\!\!\!\perp A \mid X$ (= all confounders are in $X$)




- Under Assumptions 1-3, all $\mathcal{L}_2$ causal estimands (and some $\mathcal{L}_3$) are identifiable:
  - CAPOs: $\mathbb{E}(Y[a] \mid x) = \mu_a(x)$; CATE: $\mathbb{E}(Y[1] - Y[0] \mid x) = \mu_1(x) - \mu_0(x)$
  - APOs: $\mathbb{E}(Y[a]) = \mathbb{E}[\mu_a(X)]$ (regression adjustment), $\cdots = \mathbb{E}[Y \mathbb{1}\{A = a\}/\pi_a(X)]$ (inverse probability of treatment weighting, IPTW)

---

**Estimation**

- How to estimate the functional $\psi(\mathbb{P})$ from data $\mathcal{D}$ *(finite observational data)*: $\hat{\psi}_n = \psi(\mathcal{D})$?
- Naïve approach: **plug-in estimators** ($\hat{\psi}_n^{\text{PI}}$). First, fit the nuisance functions ($\hat{\mu}_a, \hat{\pi}_a$) with flexible ML models of choice; then, **plug** them into the identification formulas:
  - CAPOs: $\hat{\mathbb{E}}(Y[a] \mid x) = \hat{\mu}_a(x)$; CATE: $\hat{\mathbb{E}}(Y[1] - Y[0] \mid x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$
  - APOs: $\widehat{\mathbb{E}(Y[a])} = \mathbb{P}_n[\hat{\mu}_a(X)], \cdots = \mathbb{P}_n[Y \mathbb{1}\{A = a\}/\hat{\pi}_a(X)]$ ($\mathbb{P}_n$ is a sample average)

# Introduction: Debiased estimation

- Causal estimands depend on **nuisance functions** $\eta$ (e.g., $\mu_a$). Plug-in estimators use learned $\hat{\eta} \Rightarrow$ the errors from $\hat{\eta}$ propagate at *first order*. E.g., for the APO, $\psi = \mathbb{E}[Y[1]]$:

**Plug-in bias**

$$\hat{\psi}_n^{\mathrm{PI}} - \psi = \underbrace{(\mathbb{P}_n - \mathbb{E})\{\mu_1(X)\}}_{\text{central limit theorem term}} + \underbrace{\mathbb{E}[\hat{\mu}_1(X) - \mu_1(X)]}_{\text{first-order nuisance error}} + o_{\mathbb{P}}(n^{-1/2})$$

- Can we remove the first-order errors of the nuisance functions?

# Introduction: Debiased estimation

**Plug-in bias**

- Causal estimands depend on **nuisance functions** $\eta$ (e.g., $\mu_a$). Plug-in estimators use learned $\hat{\eta} \Rightarrow$ the errors from $\hat{\eta}$ propagate at *first order*. E.g., for the APO, $\psi = \mathbb{E}[Y[1]]$:

$$\hat{\psi}_n^{\text{PI}} - \psi = \underbrace{(\mathbb{P}_n - \mathbb{E})\{\mu_1(X)\}}_{\text{central limit theorem term}} + \underbrace{\mathbb{E}[\hat{\mu}_1(X) - \mu_1(X)]}_{\text{first-order nuisance error}} + o_{\mathbb{P}}(n^{-1/2})$$

- Can we remove the first-order errors of the nuisance functions?

**Finite-dim. estimands**

- **One-step bias correction**: adding sample average of the efficient influence function $\varphi_a(Z; \eta)$ (augmented IPTW, A-IPTW, estimator) (Bickel, 1982; Pfanzagl, 1982):

$$\hat{\psi}_n^{\text{A-IPTW}} = \hat{\psi}_n^{\text{PI}} + \mathbb{P}_n\{\varphi_a(Z; \hat{\eta})\} = \mathbb{P}_n\left\{\frac{\mathbb{1}\{A=1\}}{\hat{\pi}_1(X)}(Y - \hat{\mu}_1(X)) + \hat{\mu}_1(X)\right\}$$

- Only second-order error term left $\Rightarrow$ semi-parametric efficiency & double robustness

# Introduction: Debiased estimation

**Plug-in bias**

- Causal estimands depend on **nuisance functions** $\eta$ (e.g., $\mu_a$). Plug-in estimators use learned $\hat{\eta} \Rightarrow$ the errors from $\hat{\eta}$ propagate at *first order*. E.g., for the APO, $\psi = \mathbb{E}[Y[1]]$:

$$\hat{\psi}_n^{\text{PI}} - \psi = \underbrace{(\mathbb{P}_n - \mathbb{E})\{\mu_1(X)\}}_{\text{central limit theorem term}} + \underbrace{\mathbb{E}[\hat{\mu}_1(X) - \mu_1(X)]}_{\text{first-order nuisance error}} + o_{\mathbb{P}}(n^{-1/2})$$

- Can we remove the first-order errors of the nuisance functions?

**Finite-dim. estimands**

- **One-step bias correction**: adding sample average of the efficient influence function $\varphi_a(Z; \eta)$ (augmented IPTW, A-IPTW, estimator) (Bickel, 1982; Pfanzagl, 1982):

$$\hat{\psi}_n^{\text{A-IPTW}} = \hat{\psi}_n^{\text{PI}} + \mathbb{P}_n\{\varphi_a(Z; \hat{\eta})\} = \mathbb{P}_n\left\{\frac{\mathbb{1}\{A=1\}}{\hat{\pi}_1(X)}(Y - \hat{\mu}_1(X)) + \hat{\mu}_1(X)\right\}$$

- Only second-order error term left $\Rightarrow$ semi-parametric efficiency & double robustness

**Functional estimands**

- The EIF is not well-defined $\Rightarrow$ **Neyman-orthogonal learning**: (i) defining the estimand as the best projection in some model class, $g^* = \arg\min_{g \in \mathcal{G}} \mathcal{L}(g, \eta)$; then, (ii) debiasing the gradient of the risk $\mathcal{L}(g, \hat{\eta})$ (Foster & Syrgkanis, 2023): $D_\eta D_g \, \mathcal{L}(g^*, \eta)[g - g^*, \hat{\eta} - \eta] = 0$
- Only second-order error term left $\Rightarrow$ quasi-oracle efficiency & double robustness

In my thesis, I develop such corrections/orthogonal risks for **distributional** and **partially identified** estimands: Part I (DPOs, CDPOs), Part III (bounds on the CDTE)

# Introduction: Three sources of uncertainty

| | **Epistemic (EU)** | | |
|---|---|---|---|
| **What?** | finite-sample uncertainty of an estimator | | |
| **Example quantifier** | $\mathrm{Var}(\hat{\psi}_n)$, confidence/credibility intervals | | |
| **Problem** | estimation | | |
| **Reducible by** | collecting more data | | |
| **Remains for $n \to \infty$?** | no | | |
| **In causal ML** | relatively well studied | | |

# Introduction: Three sources of uncertainty

| | **Epistemic (EU)** | **Aleatoric (AU)** |
|---|---|---|
| **What?** | finite-sample uncertainty of an estimator | intrinsic randomness of the causal target |
| **Example quantifier** | $\mathrm{Var}(\hat{\psi}_n)$, confidence/credibility intervals | $\mathrm{Var}(Y[a] \mid x)$, $\mathbb{F}_a(y \mid x) = \mathbb{P}(Y[a] \leq y \mid x)$ |
| **Problem** | estimation | target is random |
| **Reducible by** | collecting more data | adding predictive covariates |
| **Remains for $n \to \infty$?** | no | **yes** |
| **In causal ML** | relatively well studied | **open: Parts I & III** |

# Introduction: Three sources of uncertainty

| | **Epistemic (EU)** | **Aleatoric (AU)** | **Unobserved variables (UU)** |
|---|---|---|---|
| **What?** | finite-sample uncertainty of an estimator | intrinsic randomness of the causal target | lack of identifiability (missing variables) |
| **Example quantifier** | $\mathrm{Var}(\hat{\psi}_n)$, confidence/credibility intervals | $\mathrm{Var}(Y[a] \mid x)$, $\mathbb{F}_a(y \mid x) = \mathbb{P}(Y[a] \leq y \mid x)$ | partial identification bounds $\psi \in [\psi^L, \psi^U]$ |
| **Problem** | estimation | target is random | identification |
| **Reducible by** | collecting more data | adding predictive covariates | adding covariates / assumptions (sensitivity models) |
| **Remains for $n \to \infty$?** | no | **yes** | **yes** |
| **In causal ML** | relatively well studied | **open: Parts I & III** | **open: Parts II & III** |

# Introduction: Three sources of uncertainty

| | **Epistemic (EU)** | **Aleatoric (AU)** | **Unobserved variables (UU)** |
|---|---|---|---|
| **What?** | finite-sample uncertainty of an estimator | intrinsic randomness of the causal target | lack of identifiability (missing variables) |
| **Example quantifier** | $\mathrm{Var}(\hat{\psi}_n)$, confidence/credibility intervals | $\mathrm{Var}(Y[a] \mid x)$, $\mathbb{F}_a(y \mid x) = \mathbb{P}(Y[a] \leq y \mid x)$ | partial identification bounds $\psi \in [\psi^L, \psi^U]$ |
| **Problem** | estimation | target is random | identification |
| **Reducible by** | collecting more data | adding predictive covariates | adding covariates / assumptions (sensitivity models) |
| **Remains for $n \to \infty$?** | no | **yes** | **yes** |
| **In causal ML** | relatively well studied | **open: Parts I & III** | **open: Parts II & III** |

Traditional ML usually distinguishes only EU and AU. In causal ML, UU appears naturally, and **all sources can interact** (e.g., the distribution of the treatment effect carries both AU and UU)

# Introduction: Case study (sources of uncertainty)

**Observational data** (California's tobacco tax, Abadie et al., 2010):

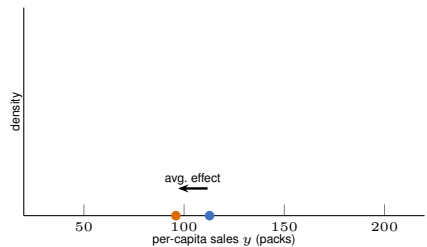
- 39 US states, 1970–2000 ($n = 1{,}209$ state-years)
- $A$: $a = 0$ (no tax); $a = 1$ ($+25$ cents tax per pack)
- $Y$: per-capita cigarette sales (packs)
- $X$: beer consumption, GDP, retail price, % aged 15–24

**Aim:** effect of the tax on cigarette sales
**Assumptions:** ✓ 1. Consistency; (✓) 2. Strong overlap; ✓ 3. Unconfoundedness

**Evidence** (here, adapted for narration purposes):

- average effect: ATE is relatively **small**

**Cigarette sales with ($a = 1$) vs. without ($a = 0$) the tax**



# Introduction: Case study (sources of uncertainty)

**Observational data** (California's tobacco tax, Abadie et al., 2010):

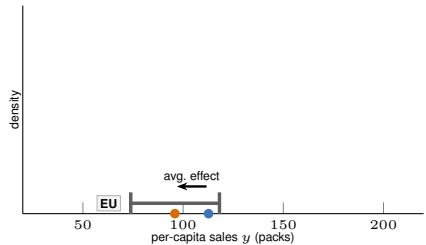
- 39 US states, 1970–2000 ($n = 1{,}209$ state-years)
- $A$: $a = 0$ (no tax); $a = 1$ ($+25$ cents tax per pack)
- $Y$: per-capita cigarette sales (packs)
- $X$: beer consumption, GDP, retail price, % aged 15–24

**Aim:** effect of the tax on cigarette sales
**Assumptions:** ✓ 1. Consistency; (✓) 2. Strong overlap; ✓ 3. Unconfoundedness

**Evidence** (here, adapted for narration purposes):

- average effect: ATE is relatively **small**

**Cigarette sales with ($a = 1$) vs. without ($a = 0$) the tax**




**EU:** only 12 treated observations ⇒ wide confidence interval

# Introduction: Case study (sources of uncertainty)

**Observational data** (California's tobacco tax, Abadie et al., 2010):

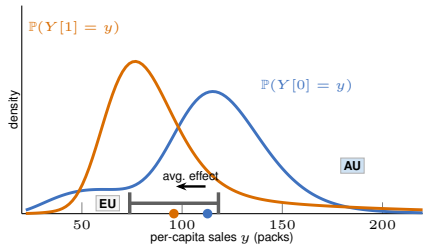
- 39 US states, 1970–2000 ($n = 1{,}209$ state-years)
- $A$: $a = 0$ (no tax); $a = 1$ ($+25$ cents tax per pack)
- $Y$: per-capita cigarette sales (packs)
- $X$: beer consumption, GDP, retail price, % aged 15–24

**Aim:** effect of the tax on cigarette sales
**Assumptions:** ✓ 1. Consistency; (✓) 2. Strong overlap; ✓ 3. Unconfoundedness

**Evidence** (here, adapted for narration purposes):

- average effect: ATE is relatively **small**
- the densities of POs: the mass of $\mathbb{P}(Y[1])$ **visibly** shifts to the left compared to $\mathbb{P}(Y[0])$

**Cigarette sales with ($a = 1$) vs. without ($a = 0$) the tax**




**EU:** only 12 treated observations ⇒ wide confidence interval

**AU:** potetial sales vary ⇒ the whole distribution of $Y[a]$ matters

# Introduction: Case study (sources of uncertainty)

**Observational data** (California's tobacco tax, Abadie et al., 2010):

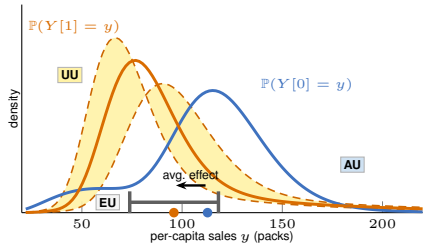
- 39 US states, 1970–2000 ($n = 1{,}209$ state-years)
- $A$: $a = 0$ (no tax); $a = 1$ ($+25$ cents tax per pack)
- $Y$: per-capita cigarette sales (packs)
- $X$: beer consumption, GDP, retail price, % aged 15–24

**Aim:** effect of the tax on cigarette sales
**Assumptions:** ✓ 1. Consistency; (✓) 2. Strong overlap;
× 3. Unconfoundedness

**Evidence** (here, adapted for narration purposes):

- average effect: ATE is relatively **small**
- the densities of POs: the mass of $\mathbb{P}(Y[1])$ **visibly** shifts to the left compared to $\mathbb{P}(Y[0])$

**Cigarette sales with ($a = 1$) vs. without ($a = 0$) the tax**

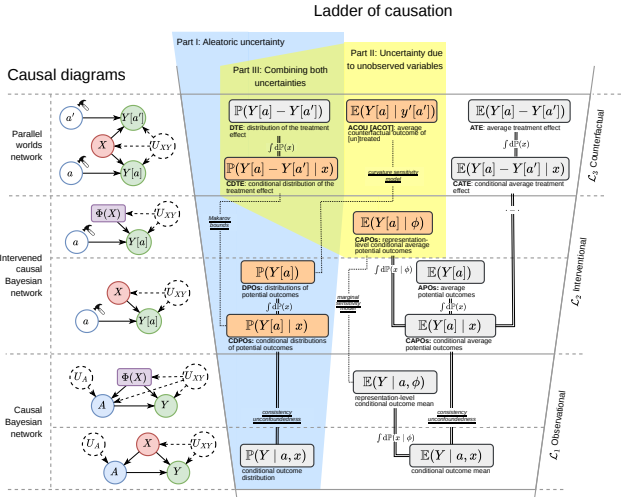



**EU:** only 12 treated observations ⇒ wide confidence interval

**AU:** potetial sales vary ⇒ the whole distribution of $Y[a]$ matters

**UU:** hidden confounding ⇒ a range of plausible densities consistent with data

# Introduction: Uncertainty-estimands map of the thesis




- Layers of causation: observational ($\mathcal{L}_1$) $\rightarrow$ interventional ($\mathcal{L}_2$) $\rightarrow$ counterfactual ($\mathcal{L}_3$)
- Double lines: **point identification**; dotted lines: **partial identification**
- **Part I** AU: distributions of potential outcomes (DPOs, CDPOs)
- **Part II** UU: representation-level CAPOs; average counterfactual outcomes
- **Part III** AU + UU: distribution of the treatment effect (CDTE)

# Agenda

**Introduction**

**Part I: AU (Aleatoric Uncertainty)**

- Normalizing Flows for Interventional Density Estimation (ICML 2023)
- GDR-learners: Orthogonal Learning of Generative Models for Potential Outcomes (ICLR 2026)

**Part III: AU + UU (Combining Both Uncertainties)**

**Summary & outlook**

"Part II: UU (Uncertainty due to Unobserved Variables)" is covered in the backup slides

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

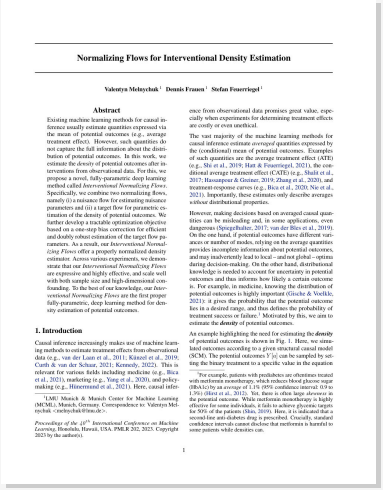
INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

PART I • ALEATORIC UNCERTAINTY

# Normalizing Flows for Interventional Density Estimation

**Valentyn Melnychuk**, Dennis Frauen, Stefan Feuerriegel

ICML 2023



Normalizing Flows for Interventional Density Estimation

Valentyn Melnychuk[1] Dennis Frauen[1] Stefan Feuerriegel[1]

Abstract

Existing machine learning methods for causal inference usually estimate quantities expressed via the mean of potential outcomes (e.g., average treatment effect). However, such quantities do not capture the full information about the distribution of potential outcomes. In this work, we estimate the density of potential outcomes after interventions from observational data. For this, we propose a novel, fully-parametric deep learning method called *Interventional Normalizing Flows*. Specifically, we combine two normalizing flows, namely (i) a nuisance flow for estimating nuisance parameters and (ii) a target flow for parametric estimation of the density of potential outcomes. We further develop a tractable optimization objective based on a one-step bias correction for efficient and doubly robust estimation of the target flow parameters. As a result, our *Interventional Normalizing Flows* offer a properly normalized density estimator. Across various experiments, we demonstrate that our *Interventional Normalizing Flows* are expressive and highly effective, and scale well with both sample size and high-dimensional confounding. To the best of our knowledge, our *Interventional Normalizing Flows* are the first proper fully-parametric, deep learning method for density estimation of potential outcomes.

1. Introduction

Causal inference increasingly makes use of machine learning methods to estimate treatment effects from observational data (e.g., van der Laan et al., 2011; Künzel et al., 2019; Curth & van der Schaar, 2021; Kennedy, 2022). This is relevant for various fields including medicine (e.g., Bica et al., 2021), marketing (e.g., Yang et al., 2020), and policy-making (e.g., Hünermund et al., 2021). Here, causal inference from observational data promises great value, especially when experiments for determining treatment effects are costly or even unethical.

The vast majority of the machine learning methods for causal inference estimate *averaged* quantities expressed by the (conditional) mean of potential outcomes. Examples of such quantities are the average treatment effect (ATE) (e.g., Shi et al., 2019; Hatt & Feuerriegel, 2021), the conditional average treatment effect (CATE) (e.g., Shalit et al., 2017; Hassanpour & Greiner, 2019; Zhang et al., 2020), and treatment response curves (e.g., Bica et al., 2020; Nie et al., 2021). Importantly, these estimates only describe averages without distributional properties.

However, making decisions based on averaged causal quantities can be misleading and, in some applications, even dangerous (Spiegelhalter, 2017; van der Bles et al., 2019). On the one hand, if potential outcomes have different variances or number of modes, relying on the average quantities provides incomplete information about potential outcomes, and may inadvertently lead to local—and not global—optima during decision-making. On the other hand, distributional knowledge is needed to account for uncertainty in potential outcomes and thus informs how likely a certain outcome is. For example, in medicine, knowing the distribution of potential outcomes is highly important (Gische & Voelkle, 2021): it gives the probability that the potential outcome lies in a desired range, and thus defines the probability of treatment success or failure.[1] Motivated by this, we aim to estimate the *density* of potential outcomes.

An example highlighting the need for estimating the *density* of potential outcomes is shown in Fig. 1. Here, we simulated outcomes according to a given structural causal model (SCM). The potential outcomes $Y[x]$ can be sampled by setting the binary treatment to a specific value in the equation

[1] LMU Munich & Munich Center for Machine Learning (MCML), Munich, Germany. Correspondence to: Valentyn Melnychuk <melnychuk@lmu.de>.

*Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

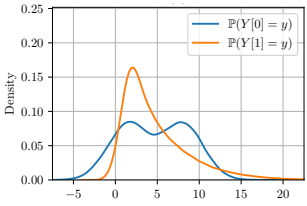
[1] For example, patients with prediabetes are oftentimes treated with metformin monotherapy, which reduces blood glucose sugar (HbA1c) by an average of 1.1% (95% confidence interval: 0.9 to 1.3%) (Hirst et al., 2012). Yet, there is often large *dissensus* in the potential outcome. While metformin monotherapy is highly effective for some individuals, it fails to achieve glycemic targets for 50% of the patients (Shin, 2019). Here, it is indicated that a second line anti-diabetic drug is prescribed. Crucially, standard confidence intervals cannot disclose that metformin is harmful to some patients while denotes can.

1

# INFs: Motivation

**From APOs to DPOs**

- Distributions of POs (DPOs), $\mathbb{P}(Y[a])$, capture the **population-level** AU: multi-modality, quantiles, probability of (un)wanted outcomes
- We call the estimation of densities of DPO an **interventional density estimation** (IDE)
- Example: same mean & variance of $Y[0]$ and $Y[1] \Rightarrow$ ATE $= 0$, but the DPOs differ: $\mathbb{P}(Y[1] < 5) \approx 0.63$ vs. $\mathbb{P}(Y[0] < 5) \approx 0.51$




**Existing IDE methods**

| Method | Parametric | Estimator | Proper density | Universal |
|---|---|---|---|---|
| Kim et al. (2018): kernel density estimation | semi- | A-IPTW | × | ✓ |
| Muandet et al. (2021): kernel mean embeddings | non- | plug-in | × | ✓ |
| Kennedy et al. (2023): exp. family / truncated series | semi- / fully | A-IPTW | ✓ / × | × / ✓ |
| **INFs (our paper)**: normalizing flows | **fully** | **A-IPTW** | ✓ | ✓ |

**Research gap:** no efficient (A-IPTW) estimator of the DPOs that is *both* a proper density *and* universal (and, hence, no fully-parametric deep learning method for IDE)

# INFs: Target parameters

**Estimand**

- Under the POF assumptions 1-3, the densities of the DPOs are identified via regression adjustment: $\mathbb{P}(Y[a] = y) = \mathbb{E}_X\left[\mathbb{P}(Y = y \mid X, A = a)\right]$
- Semi-parametric plug-in $\mathbb{P}_n\{\hat{\mathbb{P}}(Y = y \mid X, A = a)\}$ (i) is *inefficient*; (ii) *cannot regularize* the DPOs directly; & (iii) averages over the whole sample for *every* evaluation point $y$

**Target parameters**

- Parametric model $\{g(y; \beta_a)\}$ (here: a normalizing flow) with $\int g(y; \beta_a)\, \mathrm{d}y = 1$
- **Projection parameters**: best KL projection of the DPOs onto the model class
  $\beta_a^* = \arg\min_{\beta_a} \mathrm{KL}\big(\mathbb{P}(Y[a]) \,\|\, g(\cdot; \beta_a)\big) = \arg\min_{\beta_a} \mathbb{E}\big[-\log g(Y[a]; \beta_a)\big]$
- Equivalently, a moment condition with the score $T = \nabla_{\beta_a} \log g$:
  $m(\beta_a) = \mathbb{E}\big[T(Y[a]; \beta_a)\big] = 0$

**Why normalizing flows?**

- **Proper** (normalized) density: no negative or unnormalized density estimates
- **Universal** density approximators (e.g., neural spline flows)
- **Direct sampling** and fully-parametric: evaluation at test time does not depend on the sample size $n$

# INFs: From plug-in to efficient estimation

**Plug-in**

- Fit $g$ to the plug-in distribution (cross-entropy loss $\Leftrightarrow$ moment condition):

$$\hat{\beta}_a^{\text{PI}} = \arg\min_{\beta_a} \mathbb{E}_{\hat{Y} \sim \mathbb{P}_n\{\hat{\mathbb{P}}(Y \mid X, A=a)\}} \left[ -\log g(\hat{Y}; \beta_a) \right] \quad \Leftrightarrow \quad \hat{m}^{\text{PI}}(\beta_a) \overset{!}{=} 0$$

- Consistent with expressive nuisances, but neither efficient nor doubly robust

**One-step bias correction**

- Bias-corrected moment condition with the EIF (Kennedy et al., 2023):

$$\hat{m}^{\text{A-IPTW}}(\beta_a) = \hat{m}^{\text{PI}}(\beta_a) + \mathbb{P}_n\{\varphi_a(T; \hat{\mathbb{P}})\} \overset{!}{=} 0$$

$$\varphi_a(T; \mathbb{P}) = \tfrac{\mathbb{1}\{A=a\}}{\pi_a(X)} \big(T - \mathbb{E}(T \mid X, A=a)\big) + \mathbb{E}(T \mid X, A=a) - \mathbb{E}\big[\mathbb{E}(T \mid X, A=a)\big]$$

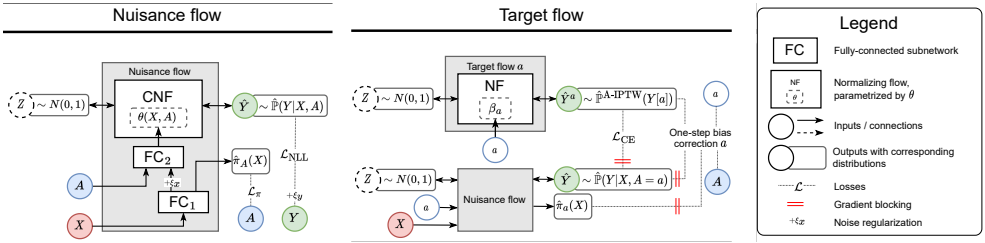
- A system of non-linear equations: hard to solve with deep learning models

**Our tractable objective**

- Taking the antiderivative wrt. $\beta_a$ yields a **debiased (Neyman-orthogonal) loss** to which we can apply gradient descent:

$$\hat{\beta}_a^{\text{A-IPTW}} = \arg\min_{\beta_a} \Big[ \underbrace{\mathbb{E}_{\hat{Y} \sim \mathbb{P}_n\{\hat{\mathbb{P}}(Y \mid X, A=a)\}} \left[ -\log g(\hat{Y}; \beta_a) \right]}_{\text{cross-entropy loss}} + \underbrace{\mathbb{P}_n \Big\{ \tfrac{\mathbb{1}\{A=a\}}{\hat{\pi}_a(X)} \Big( -\log g(Y; \beta_a) - \mathbb{E}_{Y \sim \hat{\mathbb{P}}(Y \mid X, A=a)} \left[ -\log g(Y; \beta_a) \right] \Big) \Big\}}_{\text{one-step bias correction}} \Big]$$

# INFs: Interventional Normalizing Flows

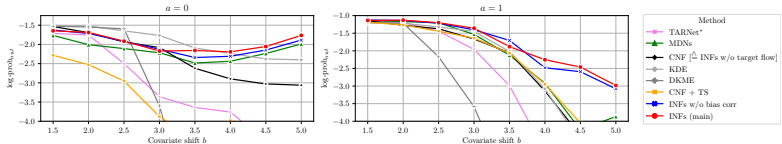



**Training**

- Stage 1: fit the nuisance flow (a conditional normalizing flow with a hypernetwork, FC$_1$ and FC$_2$) to estimate $\hat{\mathbb{P}}(Y = y \mid x, a)$ and $\hat{\pi}_a(X)$ with $\mathcal{L}_{\text{NLL}} + \alpha\,\mathcal{L}_\pi$ (negative log-likelihood + binary cross-entropy)
- Stage 2: freeze the nuisance flow & fit the target flows $g(y; \beta_a)$ (unconditional normalizing flows, one per treatment) with our debiased loss

# INFs: Experiments & takeaways

**Synthetic data:** out-sample log-prob. vs. covariate shift $b$ (larger $b$ = weaker overlap; higher is better)




**ACIC 2016 & 2018:** % of runs with best performance

| | ACIC 2016 (77 datasets) | | ACIC 2018 (24 datasets) | |
|---|---|---|---|---|
| | % $\text{best}_{\text{in}}$ | % $\text{best}_{\text{out}}$ | % $\text{best}_{\text{in}}$ | % $\text{best}_{\text{out}}$ |
| TARNet* | 3.90% | 6.23% | 7.08% | 7.50% |
| MDNs | 28.96% | 29.35% | 21.25% | 18.75% |
| CNF [≅ INFs w/o target flow] | 14.42% | 15.97% | 14.17% | 14.58% |
| KDE (Kim et al., 2018) | 1.04% | 1.04% | 10.42% | 9.58% |
| DKME (Muandet et al., 2021) | 0.39% | 0.78% | 8.75% | 10.83% |
| CNF+TS (Kennedy et al., 2023) | 8.18% | 8.96% | 5.83% | 5.42% |
| INFs w/o bias corr | 5.45% | 7.27% | 4.58% | 5.42% |
| INFs (main) | **37.66%** | **30.39%** | **27.92%** | **27.92%** |

Higher = better (best in bold)

- Based on the synthetic and semi-synthetic benchmarks, INFs outperform both neural plug-in and debiased non-neural baselines
- **Takeaways**: **INFs** are the first proper, fully-parametric deep learning method for interventional density estimation, with efficient & doubly-robust loss.

**Limitation:** So far, we estimated the distribution of POs at the population level $\rightarrow$ *modelling* individualized distributions?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN | LMU MUNICH SCHOOL OF MANAGEMENT | INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

PART I • ALEATORIC UNCERTAINTY

# GDR-learners: Orthogonal Learning of Generative Models for Potential Outcomes

**Valentyn Melnychuk**, Stefan Feuerriegel

ICLR 2026



Published as a conference paper at ICLR 2026

GDR-LEARNERS: ORTHOGONAL LEARNING OF GENERATIVE MODELS FOR POTENTIAL OUTCOMES

**Valentyn Melnychuk & Stefan Feuerriegel**
LMU Munich & Munich Center for Machine Learning
Munich, Germany
melnychuk@lmu.de

ABSTRACT

Various deep generative models have been proposed to estimate potential outcomes distributions from observational data. However, none of them have the favorable theoretical property of general Neyman-orthogonality and, associated with it, quasi-oracle efficiency and double robustness. In this paper, we introduce a general suite of generative Neyman-orthogonal (doubly-robust) learners that estimate the conditional distributions of potential outcomes. Our proposed generative doubly-robust learners (GDR-learners) are flexible and can be instantiated with many state-of-the-art deep generative models. In particular, we develop GDR-learners based on (a) conditional normalizing flows (which we call *GDR-CNFs*), (b) conditional generative adversarial networks (*GDR-CGANs*), (c) conditional variational autoencoders (*GDR-CVAEs*), and (d) conditional diffusion models (*GDR-CDMs*). Unlike the existing methods, our GDR-learners possess the properties of quasi-oracle efficiency and rate double robustness, and are thus asymptotically optimal. In a series of carrie-iyenthetic experiments, we demonstrate that our GDR-learners are very effective and outperform the existing methods in estimating the conditional distributions of potential outcomes.

1 INTRODUCTION

Causal machine learning (ML) is widely used to predict potential outcomes (POs), namely, the outcome after an intervention. In medicine, for example, an accurate prediction of POs can guide the choice of the optimal treatment from several available treatment options (Feuerriegel et al., 2024)). The POs have a central role in the causal ML as they define various causal quantities, such as treatment effects (Curth & van der Schaar, 2021) or a policy value (Qian & Murphy, 2011; Frauen et al., 2025b).

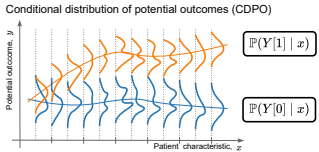
Recently, many works have suggested departing from estimating simple *conditional averages* of the POs (CAPOs) and rather aim at the *whole conditional distributions* (Louizos et al., 2017; Yoon et al., 2018; Zhang et al., 2021; Vanderschueren et al., 2023; Ma et al., 2024; Wu et al., 2025). In this way, one can capture the inherent randomness of the PO (see Fig. 1), namely, its *aleatoric uncertainty*. The aleatoric uncertainty is particularly important for reliable decision-making (Spiegelhalter, 2017; van der Bles et al., 2019): by knowing the whole distribution of the PO, decision-makers such as clinicians may be able to evaluate the probabilities of unwanted outcomes. In this work, we thus focus on learning the **conditional distributions of potential outcomes (CDPOs)**.

Various state-of-the-art generative models have been developed (or can be adapted) to model the CDPOs (Louizos et al., 2017; Kingma et al., 2019; Nemirovsky et al., 2020; Pawlowski et al., 2020; Sauer & Geiger, 2021; Zhang et al., 2021; Sanchez & Tsaftaris, 2022; Ribeiro et al., 2023; Vanderschueren et al., 2023; Ma et al., 2024; Wu et al., 2025). These works mainly differ in (a) the underlying probabilistic models (e. g., variational autoencoders (Louizos et al., 2017; Zhang et al., 2021; Ribeiro et al., 2023), diffusion models (Sanchez & Tsaftaris, 2022; Ma et al., 2024; Wu et al., 2025), etc.); and in (b) their learning objectives (e. g., plug-in (Vanderschueren et al., 2023) and inverse propensity of treatment weighted (IPTW) losses (Ma et al., 2024)).

# GDR-learners: Motivation

**From CAPOs to CDPOs**

- Conditional distributions of POs (CDPOs), $\mathbb{P}(Y[a] \mid x)$, capture the **individualized** AU: heavy tails, multi-modality, probability of unwanted outcomes
- CDPOs can be flexibly estimated with generative deep learning models




**Existing generative methods**

- Group (1) models the whole causal data-generating process
- Group (2) targets the CDPOs directly, but all are plug-in approaches

| Target | Work | Generative model | Learner type | Neyman-orthogonality: Partial | Neyman-orthogonality: General |
|---|---|---|---|---|---|
| ✗ (1) DGP | CEVAE (Louizos et al., 2017); TEDVAE (Zhang et al., 2021) | VAE | Plug-in | ✗ | ✗ |
| | CausalGAN (Kocaoglu et al., 2018); CGN (Sauer & Geiger, 2021) | GAN | Plug-in | ✗ | ✗ |
| | DSCM (Pawlowski et al., 2020) | VAE/NF | Plug-in | ✗ | ✗ |
| | Diff-SCM (Sanchez & Tsaftaris, 2022); DCM (Chao et al., 2023) | DM | Plug-in | ✗ | ✗ |
| | CausalHVAE (Ribeiro et al., 2023) | VAE | Plug-in | ✗ | ✗ |
| ✓ (2) CDPOs | GANITE (Yoon et al., 2018) | GAN | RA | ✗ | ✗ |
| | NOFLITE (Vanderschueren et al., 2023) | NF | Plug-in | ✗ | ✗ |
| | DiffPO (Ma et al., 2024) | DM | IPTW | ✓ | ✗ |
| | PO-Flow (Wu et al., 2025) | NF/DM | Plug-in | ✗ | ✗ |
| | ***GDR-learners* (our paper)** | **any** | **DR** | ✓ | ✓ |

Generative model: VAE: variational autoencoder; GAN: generative adversarial network; NF: normalizing flow; DM: diffusion model.
Learner type: RA: regression-adjusted; IPTW: inverse propensity of treatment weighted; DR: doubly-robust.

**Research gap:** no general Neyman-orthogonal learners for the CDPOs that can be instantiated with *any* deep generative model

# GDR-learners: Target risk for generative models

**Estimand**

- Under the POF assumptions 1-3, the densities of the CDPOs given covariates $V \subseteq X$ are identifiable as:

$$\mathbb{P}(Y[a] = y \mid V = v) = \mathbb{E}\big[\xi_a(y \mid X) \mid V = v\big] \qquad \text{(for } V = X\text{: } \xi_a(y \mid x)\text{)}$$

**Target risk**

- We aim to find the best distributional projection of the CDPOs onto a generative model class $\mathcal{G} = \{g_a(y, z \mid v)\}$:

$$g_a^* = \arg\max_{g_a \in \mathcal{G}} \mathcal{L}(g_a), \quad \mathcal{L}(g_a) = \mathbb{E}\Big[\mathbb{E}_{Z \sim \varepsilon_z} \log g_a(Y[a], Z \mid V)\Big]$$

- The risk can be identified as

$$\mathcal{L}(g_a) = \mathbb{E}\left[\int \mathbb{E}_Z \log g_a(y, Z \mid V)\, \xi_a(y \mid X)\, \mathrm{d}y\right] = \mathbb{E}\left[\frac{\mathbb{1}\{A=a\}}{\pi_a(X)} \mathbb{E}_Z \log g_a(Y, Z \mid V)\right]$$

**Covers many generative models**

| Generative model | $g_a(y, z \mid v)$ | Projection wrt. |
|---|---|---|
| (a) conditional normalizing flow (CNF) | $p_a(y \mid v)$ | KL divergence |
| (b) conditional GAN (CGAN) | $d_a(y \mid v) \cdot \big(1 - d_a(f_a(z \mid v) \mid v)\big)$ | Jensen–Shannon div. |
| (c) conditional VAE (CVAE) | $p_a(y, z \mid v) \,/\, q_a(z \mid y, v)$ | KL + inference gap |
| (d) conditional diffusion model (CDM) | $p_a(y, z \mid v) \,/\, q_a(z \mid y)$ | KL + inference gap |

# GDR-learners: From plug-in to doubly robust learners

**Plug-in learner**

- $\hat{\mathcal{L}}_{\text{PI}}(g_a) = \mathbb{P}_n \Big\{ \mathbb{1}\{A = a\} \, \mathbb{E}_Z \log g_a(Y, Z \mid X) \Big\}$

$\rightarrow$ (i) fits the CDPOs well only on factual data (= covariate shift); (ii) is *inefficient*; & (iii) only works for the full covariates set $X$

**Two-stage learners: RA & IPTW**

- $\hat{\mathcal{L}}_{\text{RA}}(g_a, \hat{\xi}_a) = \mathbb{P}_n \Big\{ \mathbb{1}\{A = a\} \mathbb{E}_Z \log g_a(Y, Z \mid V) + \mathbb{1}\{A \neq a\} \int \mathbb{E}_Z \log g_a(y, Z \mid V) \, \hat{\xi}_a(y \mid X) \, \mathrm{d}y \Big\}$
- $\hat{\mathcal{L}}_{\text{IPTW}}(g_a, \hat{\pi}_a) = \mathbb{P}_n \left\{ \frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a(X)} \, \mathbb{E}_Z \log g_a(Y, Z \mid V) \right\}$

$\rightarrow$ both estimate the right target risk, but nuisance errors propagate at *first order*

**GDR-learner (ours)**

- One-step bias correction of the RA-learner, with the nuisance functions $\hat{\eta} = (\hat{\xi}_a, \hat{\pi}_a)$:

$$\hat{\mathcal{L}}_{\text{GDR}}(g_a, \hat{\eta}) = \mathbb{P}_n \Big\{ \frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a(X)} \, \mathbb{E}_Z \log g_a(Y, Z \mid V) + \left( 1 - \frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a(X)} \right) \int \mathbb{E}_Z \log g_a(y, Z \mid V) \, \hat{\xi}_a(y \mid X) \, \mathrm{d}y \Big\}$$

- We approximate the integral via Monte Carlo samples from $\hat{\xi}_a \Rightarrow$ the nuisance model only needs to *sample* (w/o explicit density)

# GDR-learners: Theoretical guarantees

**Theorem 1: Neyman-orthogonality**

The risk of GDR-learners is Neyman-orthogonal:

$$D_\eta D_g \, \mathcal{L}_{\mathrm{GDR}}(g_a^*, \eta)[g_a - g_a^*, \hat{\eta} - \eta] = 0 \quad \text{for all } g_a \in \mathcal{G} \text{ and } \hat{\eta} \in \mathcal{H}$$

$\rightarrow$ the risk is first-order insensitive to errors in the nuisance functions

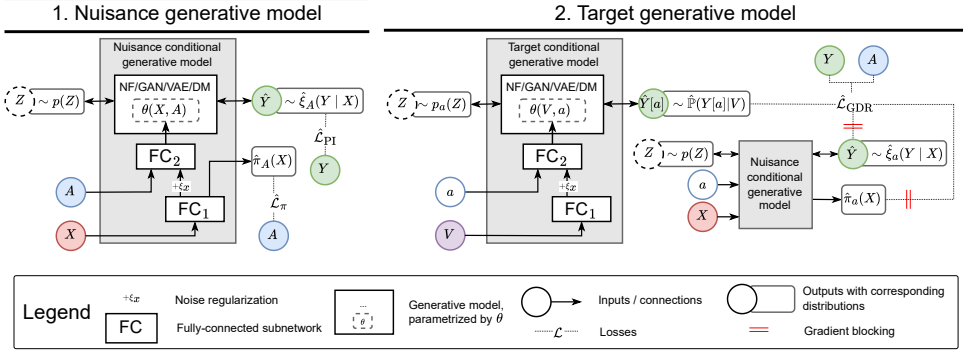
**Theorem 2: quasi-oracle efficiency & double robustness**

Under mild model-dependent convexity conditions:

$$\|g_a^* - \hat{g}_a\|_{\mathcal{G}}^2 \lesssim \underbrace{\mathcal{L}_{\mathrm{GDR}}(\hat{g}_a, \hat{\eta}) - \mathcal{L}_{\mathrm{GDR}}(g_a^*, \hat{\eta})}_{\text{optimization error}} + \underbrace{\|\xi_a - \hat{\xi}_a\|_{L_4}^2 \, \|\pi_a - \hat{\pi}_a\|_{L_4}^2}_{\text{higher-order nuisance error}}$$

- **Quasi-oracle efficiency:** when the nuisance functions are estimated at the rate $o_{\mathbb{P}}(n^{-1/4})$, learning the target model with the debiased risk is as if they were known
- **Rate double robustness:** if one nuisance function converges slowly, the final error can be compensated by a faster convergence of the other nuisance

# GDR-learners: Instantiations with deep generative models




**Four instantiations**

**GDR-CNFs** (normalizing flows) • **GDR-CGANs** (GANs) • **GDR-CVAEs** (VAEs) • **GDR-CDMs** (diffusion models)

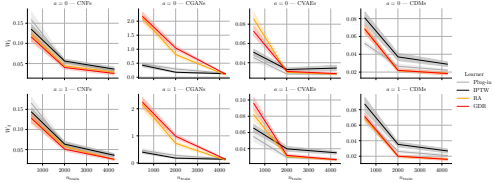
**Training**

Stage 1: fit the nuisance generative model to estimate $\hat{\xi}_a(y \mid x)$ with $\hat{\mathcal{L}}_{\text{PI}}$ + binary cross-entropy loss for $\hat{\pi}_a$

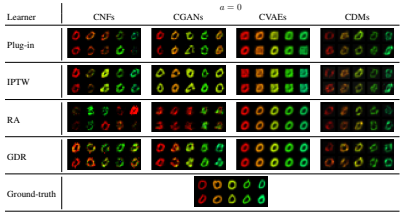
Stage 2: freeze Stage 1 model & fit the target generative models $g_0, g_1$ with $\hat{\mathcal{L}}_{\text{GDR}}$

# GDR-learners: Experiments & takeaways

**Synthetic data:** out-sample $W_2$ vs. $n$ (lower is better)




**Colored MNIST:** samples of $Y[0]$ (high-dim. outcomes)




- 16 models (4 generative models $\times$ 4 learners), covering existing baselines (e.g., GANITE $\approx$ RA-CGAN, NOFLITE $\approx$ plug-in CNF, DiffPO $\approx$ IPTW-CDM)
- GDR-learners outperform other methods as $n$ grows (as expected from the asymptotic theory)
- **Takeaways: GDR-learners** turn *any* conditional generative model into a Neyman-orthogonal learner of the CDPOs (with the properties of quasi-oracle efficiency and double robustness)

**Limitation:** Can we model the individualized distributions of the treatment effects? $\rightarrow$ Part III

# Agenda

**Introduction**

**Part I: AU (Aleatoric Uncertainty)**

**Part III: AU + UU (Combining Both Uncertainties)**

- Quantifying Aleatoric Uncertainty of the Treatment Effect: A Novel Orthogonal Learner (NeurIPS 2024)

**Summary & outlook**

"Part II: UU (Uncertainty due to Unobserved Variables)" is covered in the backup slides

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN | LMU MUNICH SCHOOL OF MANAGEMENT | INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

PART III • COMBINING BOTH UNCERTAINTIES

# Quantifying Aleatoric Uncertainty of the Treatment Effect: A Novel Orthogonal Learner

**Valentyn Melnychuk**, Stefan Feuerriegel, Mihaela van der Schaar

NeurIPS 2024



Quantifying Aleatoric Uncertainty of the Treatment Effect: A Novel Orthogonal Learner

Valentyn Melnychuk[1,2], Stefan Feuerriegel[1], Mihaela van der Schaar[2]
[1]LMU Munich & Munich Center for Machine Learning (MCML), Germany
[2]University of Cambridge & Alan Turing Institute, United Kingdom
[1]Correspondence: melnychuk@lmu.de

**Abstract**

Estimating causal quantities from observational data is crucial for understanding the safety and effectiveness of medical treatments. However, to make reliable inferences, medical practitioners require not only estimating averaged causal quantities, such as the conditional average treatment effect, but also understanding the randomness of the treatment effect as a random variable. This randomness is referred to as *aleatoric uncertainty* and is necessary for understanding the probability of benefit from treatment or quantiles of the treatment effect. Yet, the aleatoric uncertainty of the treatment effect has received surprisingly little attention in the causal machine learning community. To fill this gap, we aim to quantify the aleatoric uncertainty of the treatment effect at the covariate-conditional level, namely, the conditional distribution of the treatment effect (CDTE). Unlike average causal quantities, the CDTE is *not* point identifiable without strong additional assumptions. As a remedy, we employ partial identification to obtain sharp bounds on the CDTE and thereby quantify the aleatoric uncertainty of the treatment effect. We then develop a novel, orthogonal learner for the bounds on the CDTE, which we call AU-learner. We further show that our AU-learner has several strengths in that it satisfies Neyman-orthogonality and, thus, quasi-oracle efficiency. Finally, we propose a fully-parametric deep learning instantiation of our AU-learner.

## 1 Introduction

Estimating causal quantities from observational data is crucial for decision-making in medicine [9, 12, 22, 30, 70]. For example, medical practitioners are interested in estimating the effect of chemotherapy vs. immunotherapy on patient survival from electronic health records to understand the best treatment strategies in cancer care. Here, common estimation targets are *averaged* causal quantities such as the conditional average treatment effect (CATE) or average treatment effect (ATE). However, an averaged causal quantities do not allow for understanding the variability of the treatment effect.

What is needed for the reliability of causal quantities in medicine? To obtain *reliable* causal quantities, one often needs to "move beyond the mean" [44, 68] and consider the inherent randomness in the treatment effect as a random variable. This randomness is referred to as *aleatoric uncertainty* [17, 60, 110]. Quantifying the aleatoric uncertainty of the treatment effect is relevant in medical practice to understand the probability of benefit from treatment [26, 68] and the quantiles and variance of the treatment effect [5, 17, 26, 33, 59]. As an example, averaged quantities such as the CATE would simply suggest a positive effect for some patients, while the probability of benefit from treatment can inform patients about the odds of being negatively affected by the treatment. Hence, aleatoric uncertainty of the treatment effect promises additional, fine-grained insights beyond simple averages.

Methods for quantifying the aleatoric uncertainty of the treatment effect have gained surprisingly little attention in the causal machine learning community. So far, machine learning for treatment

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

# AU-learner: Motivation

**AU of the treatment effect**

- The treatment effect $Y[1] - Y[0]$ is itself a **random variable** with the associated conditional distribution (CDTE):

$$\mathbb{F}(\delta \mid x) = \mathbb{P}(Y[1] - Y[0] \leq \delta \mid x) \qquad \mathbb{F}^{-1}(\alpha \mid x) = \inf\{\delta : \alpha \leq \mathbb{F}(\delta \mid x)\}$$

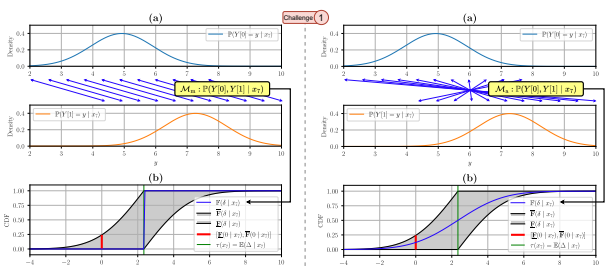
- While CATE can be "positive on average"; CDTE tells the probability of the treatment benefit or harm ($\delta = 0$), quantiles and variance of the effect

---

CDTE is **not** identifiable with Assumptions 1-3 of the POF (or even RCTs):

**Challenge 1: not point identifiable under the POF**

- Two models (monotone vs. antitone coupling) with the **same** CDPOs $\mathbb{P}(Y[a] \mid x)$ but **different** joints $\mathbb{P}(Y[0], Y[1] \mid x) \Rightarrow$ different CDTEs
- Point identification needs very strong assumptions (e.g., rank invariance / bijectivity of latent noise)



# AU-learner: Partial identification with Makarov bounds

**Makarov bounds**

- Sharp, assumption-free bounds via sup/inf convolutions of the conditional outcome CDFs $\mathbb{F}_a(y \mid x)$ (Makarov, 1982; Fan & Park, 2010):

$$\underline{\mathbb{F}}(\delta \mid x) = \left[ \sup_y \left\{ \mathbb{F}_1(y \mid x) - \mathbb{F}_0(y - \delta \mid x) \right\} \right]_+$$

$$\overline{\mathbb{F}}(\delta \mid x) = 1 + \left[ \inf_y \left\{ \mathbb{F}_1(y \mid x) - \mathbb{F}_0(y - \delta \mid x) \right\} \right]_-$$

Pointwise sharp: $\underline{\mathbb{F}} \leq \mathbb{F} \leq \overline{\mathbb{F}}$
(analogous bounds for quantiles)

**Challenge 2: no closed form**

- Unlike the CATE, Makarov bounds have no closed form in the nuisance functions

$\Rightarrow$ CATE learners do not carry over, and no efficient influence function (EIF) is known

**Challenge 3: constrained target**

- Unlike the CATE, Makarov bounds are monotone in $\delta$ and contained in $[0, 1]$

$\Rightarrow$ learning targets of debiased risks may violate these constraints (especially for small samples)

**Research gap:** no Neyman-orthogonal learners for Makarov bounds on the CDTE (existing approaches use only plug-in estimators)

# AU-learner: Learners for Makarov bounds

**Plug-in learner**

- $\hat{\underline{\mathbb{F}}}_{\mathrm{PI}}(\delta \mid x) = \left[\sup_y \{\hat{\mathbb{F}}_1(y \mid x) - \hat{\mathbb{F}}_0(y - \delta \mid x)\}\right]_+$

$\rightarrow$ (i) $\hat{\mathbb{F}}_a$ is only learned well on factual data (= covariate shift); (ii) is *inefficient*; & (iii) does not target the bounds, which are often *less heterogeneous* than the CDPOs

**Target risk**

- Projecting Makarov bounds on the target class model $g(\delta, x) \in \mathcal{G}$ (non-decreasing in $\delta$, values in $[0, 1]$) with the continuous ranked probability score (CRPS) (or squared Wasserstein risk $W_2^2$ for quantiles):

$$\mathcal{L}_{\mathrm{CRPS}}(g) = \mathbb{E}\left[\int \left(\underline{\mathbb{F}}(\delta \mid X) - g(\delta, X)\right)^2 \mathrm{d}\delta\right]$$

**RA-learner**

- $\hat{\mathcal{L}}_{\mathrm{PI}}(g, \hat{\eta}) = \mathbb{P}_n\left\{\int \left(\hat{\underline{\mathbb{F}}}_{\mathrm{PI}}(\delta \mid X; \hat{\eta}) - g(\delta, X)\right)^2 \mathrm{d}\delta\right\}$

$\rightarrow$ targets the bounds directly, but is *sensitive to nuisance function errors*

**AU-learner (ours)**

- $\gamma$-scaled one-step bias correction with the EIF term $\underline{C}$ (next slide) and nuisance functions $\hat{\eta} = (\hat{\pi}, \hat{\mathbb{F}}_0, \hat{\mathbb{F}}_1)$:

$$\hat{\mathcal{L}}_{\mathrm{AU}}(g, \hat{\eta}) = \mathbb{P}_n\left\{\int \Big(\underbrace{\hat{\underline{\mathbb{F}}}_{\mathrm{PI}}(\delta \mid X; \hat{\eta}) + \gamma\, \underline{C}(\delta, Z; \hat{\eta})}_{\text{pseudo-CDF}} - g(\delta, X)\Big)^2 \mathrm{d}\delta\right\}$$

# AU-learner: Theoretical results

**Theorem 1: EIF of Makarov bounds**

- Under mild conditions (finite argmax sets + CDFs are continuously differentiable + margin condition), averaged Makarov bounds are pathwise differentiable with the EIF:

$$\varphi\big(\mathbb{E}[\underline{\mathbb{F}}(\delta \mid X)]; \mathbb{P}\big) = \underline{C}(\delta, Z; \eta) + \underline{\mathbb{F}}(\delta \mid X; \eta) - \mathbb{E}\big[\underline{\mathbb{F}}(\delta \mid X; \eta)\big], \quad \text{where}$$

$$\underline{C}(\delta, Z; \eta) = I(X; \eta) \Big[ \tfrac{A}{\pi(X)} \big( \mathbb{1}\{Y \leq y^*\} - \mathbb{F}_1(y^* \mid X) \big) - \tfrac{1-A}{1-\pi(X)} \big( \mathbb{1}\{Y \leq y^* - \delta\} - \mathbb{F}_0(y^* - \delta \mid X) \big) \Big]$$

- $y^* \in \arg\max_y \{\mathbb{F}_1(y \mid X) - \mathbb{F}_0(y - \delta \mid X)\}$ and $I(X; \eta) = \mathbb{1}\{\sup_y \{\mathbb{F}_1(y \mid X) - \mathbb{F}_0(y - \delta \mid X)\} > 0\}$

$\rightarrow$ resolves Challenge 2

**Theorem 2**

- For $\gamma = 1$, the AU-learner risks are **Neyman-orthogonal** and achieve **quasi-oracle efficiency** whenever the nuisance functions are estimated at rate $o_{\mathbb{P}}(n^{-1/4})$

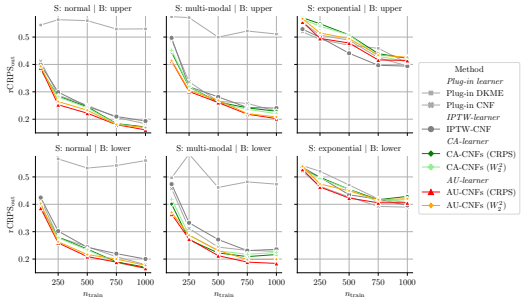
**Scaling $\gamma$**

- Pseudo-CDFs may leave $[0, 1]$ or be non-monotone
- $\gamma \in (0, 1]$ interpolates between the RA-learner ($\gamma = 0$, valid pseudo-CDFs) and the orthogonal AU-learner ($\gamma = 1$), akin to a Newton–Raphson step size

$\rightarrow$ addresses Challenge 3

# AU-learner: Instantiation, experiments & takeaways

**Synthetic data:** out-sample rCRPS vs. $n_{\text{train}}$ (lower is better)




**HC-MNIST** ($d_x = 785$, lower is better)

| | B: upper | |
|---|---|---|
| | $\text{rCRPS}_{\text{out}}$ | $W_{2\,\text{out}}$ |
| Plug-in CNF | $0.399 \pm 0.162$ | $1.051 \pm 0.514$ |
| IPTW-CNF | $0.385 \pm 0.501$ | $0.986 \pm 1.936$ |
| CA-CNFs (CRPS) | $0.382 \pm 0.045$ | $1.029 \pm 0.119$ |
| CA-CNFs ($W_2^2$) | $0.494 \pm 0.229$ | $1.239 \pm 0.541$ |
| AU-CNFs (CRPS) | $\mathbf{0.347 \pm 0.114}$ | $\mathbf{0.940 \pm 0.389}$ |
| AU-CNFs ($W_2^2$) | $0.422 \pm 0.189$ | $1.065 \pm 0.523$ |

| | B: lower | |
|---|---|---|
| | $\text{rCRPS}_{\text{out}}$ | $W_{2\,\text{out}}$ |
| Plug-in CNF | $0.391 \pm 0.104$ | $1.003 \pm 0.333$ |
| IPTW-CNF | $0.447 \pm 0.218$ | $1.141 \pm 0.917$ |
| CA-CNFs (CRPS) | $0.388 \pm 0.051$ | $1.011 \pm 0.146$ |
| CA-CNFs ($W_2^2$) | $0.446 \pm 0.148$ | $1.089 \pm 0.341$ |
| AU-CNFs (CRPS) | $\mathbf{0.357 \pm 0.097}$ | $\mathbf{0.948 \pm 0.344}$ |
| AU-CNFs ($W_2^2$) | $0.415 \pm 0.190$ | $1.025 \pm 0.520$ |

Lower = better (best in bold)

- We instantiate AU-learner with conditional normalizing flows (CNFs). Training is in two stages (analogous to Part I): nuisance CNF for $(\hat{\pi}, \hat{\mathbb{F}}_0, \hat{\mathbb{F}}_1)$ + two target CNFs (lower & upper bound)
- AU-CNF outperforms other meta-learners based on the synthetic / semi-synthetic benchmarks
- **Takeaways: AU-learner** is the first orthogonal learner for covariate-conditional Makarov bounds: it quantifies the **AU of the treatment effect** under **UU** (non-identifiability)

# Agenda

**Introduction**

**Part I: AU (Aleatoric Uncertainty)**

**Part III: AU + UU (Combining Both Uncertainties)**

**Summary & outlook**

- Contributions & takeaways
- Future directions

"Part II: UU (Uncertainty due to Unobserved Variables)" is covered in the backup slides

# Summary: Contributions of this thesis

| Part | Paper | Causal estimand | Uncert. | Key contribution |
|---|---|---|---|---|
| I | INFs<br>ICML 2023 | DPOs $\mathbb{P}(Y[a])$ | AU | First proper, fully-parametric deep learning method for interventional density estimation; efficient & doubly robust training objective |
| I | GDR-learners<br>ICLR 2026 | CDPOs $\mathbb{P}(Y[a] \mid x)$ | AU | Neyman-orthogonal learners for *any* conditional generative model (CNFs, CGANs, CVAEs, CDMs); quasi-oracle efficient & doubly robust |
| II | RICB bounds<br>ICLR 2024, spotlight | representation-level CATE $\mathbb{E}(Y[1] - Y[0] \mid \Phi = \phi)$ | UU | Formalizes representation-induced confounding bias; representation-agnostic refutation framework with bounds |
| II | Curvature Sensitivity Model<br>NeurIPS 2023, spotlight | ECOU / ECOT $\mathbb{E}(Y[a] \mid a', y')$ | UU | Non-informativeness result; first partial identification (sensitivity) model for Markovian SCMs with continuous outcomes |
| III | AU-learner<br>NeurIPS 2024 | Makarov bounds on the CDTE $\mathbb{P}(Y[1] - Y[0] \leq \delta \mid x)$ | AU + UU | EIF of Makarov bounds; first Neyman-orthogonal learner for covariate-conditional bounds on the distribution of the treatment effect |

**Common recipe:** *identification* first (point or partial, with sensitivity models where needed), then *debiased estimation* (one-step bias correction / Neyman-orthogonal learners) with flexible deep learning models

# Summary: Takeaways & outlook

| | |
|---|---|
| **Main message** | Uncertainty in causal ML is **not a single object**: finite data (EU), intrinsic randomness (AU), and limited identifiability (UU) require *different* estimands, assumptions, and estimators |
| **Practical considerations** | • **Match estimands to decisions**: means vs. distributions (tail risks, probability of success) vs. bounds<br>• **Assumption transparency**: state assumptions, diagnose them, and vary them via sensitivity analysis<br>• **Communicate the type of uncertainty**: confidence/credibility intervals (EU), full distribution of a causal target (AU), partial identification bounds (UU) |
| **Future directions** | • Longitudinal / dynamic treatment regimes; multi-valued and continuous treatments<br>• High-dimensional outcomes (images, text, multivariate clinical measurements); multiple data sources<br>• Unified, decision-theoretic integration of EU + AU + UU<br>• Evaluation & model selection for distributional and partially identified targets |

# Acknowledgments

I am deeply grateful to:

| | |
|---|---|
| **My coauthors & supervisors** | Dennis Frauen, Jonas Schweisthal, Emil Javurek, Konstantin Hess, Maresa Schröder, Haorui Ma, Yuchen Ma, Marie Brockschmidt, Vahid Balazadeh, Frederic Saab, Prof. Dr. Rahul G. Krishnan, Prof. Dr. Mihaela van der Schaar, Prof. Dr. Stefan Feuerriegel, for the ideas we chased together, the deadlines we survived together, and for making the lab feel like a second home |
| **My friends & parents** | for their support and encouragement throughout the PhD |
| **My partner** | Frederic, for the patience, care, and support |

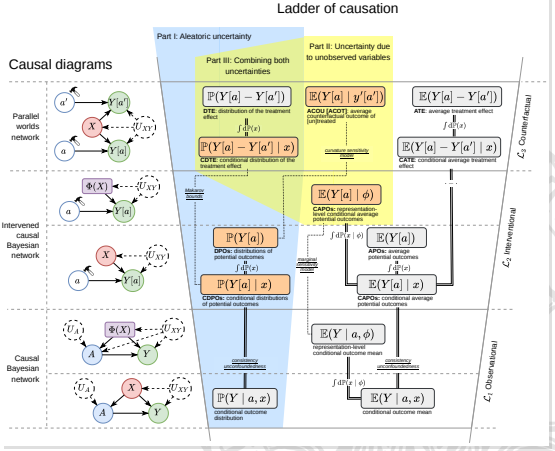


# Takeaways

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 the estimand, assumptions, and estimator.

**Thank you for your attention!**

Questions?




Code of all papers: github.com/Valentyn1997

# References I

## Papers of this dissertation

**Melnychuk**, **V.** (2026). Uncertainty Quantification in Causal Machine Learning. PhD thesis. LMU Munich.

**Melnychuk**, **V.** and S. Feuerriegel (2026). GDR-learners: Orthogonal Learning of Generative Models for Potential Outcomes. In: *International Conference on Learning Representations (ICLR)*.

**Melnychuk**, **V.**, S. Feuerriegel, and M. van der Schaar (2024a). Quantifying Aleatoric Uncertainty of the Treatment Effect: A Novel Orthogonal Learner. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 37, pp. 105039–105089.

**Melnychuk**, **V.**, D. Frauen, and S. Feuerriegel (2023a). Normalizing Flows for Interventional Density Estimation. In: *International Conference on Machine Learning (ICML)*. Vol. 202. Proceedings of Machine Learning Research, pp. 24361–24397.

**Melnychuk**, **V.**, D. Frauen, and S. Feuerriegel (2023b). Partial Counterfactual Identification of Continuous Outcomes with a Curvature Sensitivity Model. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 36, pp. 32020–32060.

**Melnychuk**, **V.**, D. Frauen, and S. Feuerriegel (2024b). Bounds on Representation-Induced Confounding Bias for Treatment Effect Estimation. In: *International Conference on Learning Representations (ICLR)*.

## Other references

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. In: *Journal of the American Statistical Association* 105.490, pp. 493–505.

Alaa, A. M., Z. Ahmad, and M. van der Laan (2023). Conformal Meta-learners for Predictive Inference of Individual Treatment Effects. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 36, pp. 47682–47703.

Balakrishnan, S., E. H. Kennedy, and L. Wasserman (2026). The Fundamental Limits of Structure-Agnostic Functional Estimation. In: *Statistical Science* 41.3, pp. 659–670.

# References II


Bickel, P. J. (1982). On Adaptive Estimation. In: *The Annals of Statistics* 10.3, pp. 647–671.

Biddinger, K. J., C. A. Emdin, M. E. Haas, et al. (2022). Association of Habitual Alcohol Intake With Risk of Cardiovascular Disease. In: *JAMA Network Open* 5.3, e223849.

Di Castelnuovo, A., S. Costanzo, V. Bagnardi, M. B. Donati, L. Iacoviello, and G. de Gaetano (2006). Alcohol Dosing and Total Mortality in Men and Women: An Updated Meta-analysis of 34 Prospective Studies. In: *Archives of Internal Medicine* 166.22, pp. 2437–2445.

Fan, Y. and S. S. Park (2010). Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference. In: *Econometric Theory* 26.3, pp. 931–951.

Foster, D. J. and V. Syrgkanis (2023). Orthogonal Statistical Learning. In: *The Annals of Statistics* 51.3, pp. 879–908.

Jin, J. and V. Syrgkanis (2025). Structure-agnostic Optimality of Doubly Robust Learning for Treatment Effect Estimation (Extended Abstract). In: *Conference on Learning Theory (COLT)*. Vol. 291. Proceedings of Machine Learning Research, pp. 3159–3160.

Kennedy, E. H., S. Balakrishnan, and L. A. Wasserman (2023). Semiparametric Counterfactual Density Estimation. In: *Biometrika* 110.4, pp. 875–896.

Kim, K., J. Kim, and E. H. Kennedy (2018). Causal Effects Based on Distributional Distances. arXiv preprint arXiv:1806.02935.

Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. New York: Springer.

Lei, L. and E. J. Candès (2021). Conformal Inference of Counterfactuals and Individual Treatment Effects. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.5, pp. 911–938.

Makarov, G. D. (1982). Estimates for the Distribution Function of a Sum of Two Random Variables When the Marginal Distributions are Fixed. In: *Theory of Probability and Its Applications* 26.4, pp. 803–806.

**Melnychuk**, **V.**, D. Frauen, J. Schweisthal, and S. Feuerriegel (2026). Orthogonal Representation Learning for Estimating Causal Quantities. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 300. Proceedings of Machine Learning Research, pp. 5086–5094.

# References III


Muandet, K., M. Kanagawa, S. Saengkyongam, and S. Marukatat (2021). Counterfactual Mean Embeddings. In: *Journal of Machine Learning Research* 22.162, pp. 1–71.

Nie, X. and S. Wager (2021). Quasi-Oracle Estimation of Heterogeneous Treatment Effects. In: *Biometrika* 108.2, pp. 299–319.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press.

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. Lecture Notes in Statistics 13. New York: Springer.

Ray, K. and A. van der Vaart (2020). Semiparametric Bayesian Causal Inference. In: *The Annals of Statistics* 48.5, pp. 2999–3020.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Schulte, R., D. Rügamer, and T. Nagler (2025). Adjustment for Confounding using Pre-Trained Representations. In: *International Conference on Machine Learning (ICML)*. Vol. 267. Proceedings of Machine Learning Research, pp. 53557–53580.

Stone, C. J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. In: *The Annals of Statistics* 10.4, pp. 1040–1053.

The New York Times (2018). Major Study of Drinking Will Be Shut Down. News article, 15 June 2018.

Yiu, A., E. Fong, C. Holmes, and J. Rousseau (2025). Semiparametric Posterior Corrections. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 87.4, pp. 1025–1054.

Zhao, J., T. Stockwell, T. Naimi, S. Churchill, J. Clay, and A. Sherk (2023). Association Between Daily Alcohol Intake and Risk of All-Cause Mortality: A Systematic Review and Meta-analyses. In: *JAMA Network Open* 6.3, e236185.

# Backup slides

### Part II: Uncertainty due to unobserved variables

- RICB: Representation-induced confounding bias
- RICB: Neural refutation framework
- CSM: Non-informative counterfactual bounds
- CSM: Curvature Sensitivity Model & APID

### Background

- Epistemic uncertainty: efficiency, optimality, finite samples
- Neyman-orthogonality, quasi-oracle efficiency, double robustness
- Overview of approaches for EU / AU / UU

### Part I & III: details

- INFs as a special case of GDR-learners
- INFs: design choices & efficiency
- GDR-learners: are $o_{\mathbb{P}}(n^{-1/4})$ rates achievable?
- AU-learner: sharpness, assumptions & low-sample regime
- AU-learner vs. conformal prediction / Bayesian methods

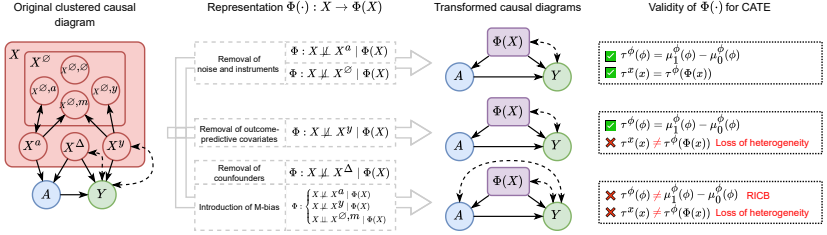
### Other

- Limitations of this thesis
- Full publication list (2025 – 2026)
- Full publication list (2022 – 2024)
- References

# Backup | Part II: Representation-induced confounding bias

**Setting**

- Representation learning for CATE (e.g., TARNet, CFR, BNN) reduces the variance of low-sample estimation via a low-dimensional (balanced/invertible) representation $\Phi(X)$




**Problem**

- If $\Phi(X)$ loses information about confounders, unconfoundedness wrt. $\Phi(X)$ fails: $\mathbb{E}(Y[a] \mid \phi) \neq \mathbb{E}(Y \mid A = a, \Phi(X) = \phi)$: **representation-induced confounding bias (RICB)** (Melnychuk et al., 2024b)

$\rightarrow$ CATE is non-identifiable from $\mathbb{P}(\Phi(X), A, Y)$, in addition to a loss of heterogeneity

← backup overview

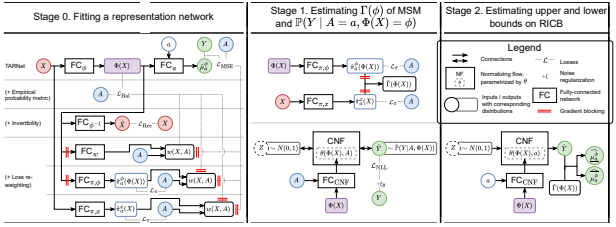
# Backup | Part II: Neural refutation framework for the RICB

**Partial identification**

- Marginal sensitivity model with a *representation-dependent* parameter $\Gamma(\phi) \geq 1$:

$$\Gamma(\phi)^{-1} \leq \frac{\pi_0^{\phi}(\phi)}{\pi_1^{\phi}(\phi)} \cdot \frac{\pi_1^{x}(x)}{\pi_0^{x}(x)} \leq \Gamma(\phi) \quad \text{for all } x\text{: } \Phi(x) = \phi$$

- No unobserved confounders: $\Gamma(\phi)$ is **estimable** from $\mathbb{P}(X, \Phi(X), A, Y)$ (unlike classical sensitivity analysis); $\Gamma \equiv 1$: no information lost

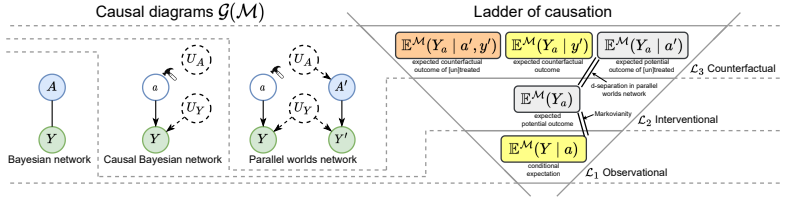



**Use & results**

- Works with any representation learning method
- Policies with deferral (treat if lower bound $> 0$, do nothing if upper bound $< 0$, else defer) reduce error rates on non-deferred decisions

← backup overview

# Backup | Part II: Partial counterfactual identification (NeurIPS 2023)

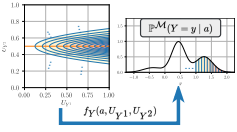



**Query**

- Expected counterfactual outcome of [un]treated (ECOU/ECOT) in Markovian SCMs with continuous outcomes: $Q_{a' \to a}(y') = \mathbb{E}(Y_a \mid a', y')$
- A continuous analogue of probabilities of necessity & sufficiency

**Theorem 1**

- Point identification needs strong assumptions (e.g., bijective generative mechanisms, BGMs)
- Without them, already for continuously differentiable SCMs $\mathfrak{B}(C^1, d)$, $d \geq 2$, the ignorance interval of ECOU [ECOT] is **non-informative** (Melnychuk et al., 2023b), i.e., the full support $[l_a, u_a]$
- Intuition: “bend” the counterfactual level sets around the factual one



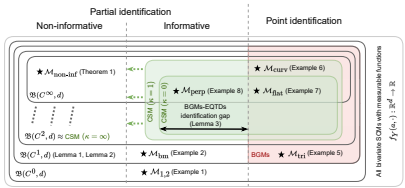

← backup overview

# Backup | Part II: Curvature Sensitivity Model & APID

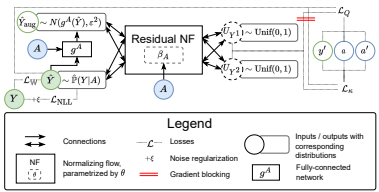
**Curvature Sensitivity Model**

- Bound the principal curvatures $\kappa_i$ of the level sets $E(y,a)$ of $f_Y(a,\cdot)$ (class $\mathfrak{B}(C^2,d)$):
  $$\kappa = \max_{a,\, y,\, u_Y \in E(y,a)} \; \max_i \; |\kappa_i(u_Y)|$$
- **Theorem 2:** under the CSM, the bounds on ECOU [ECOT] are **informative**, $l(\kappa,d) > l_a$ and $u(\kappa,d) < u_a$
- $\kappa$ spans the spectrum from BGM-type identification ($\kappa = 0$) to non-informative bounds ($\kappa \to \infty$)

**SCM classes: partial vs. point identification**




**Augmented Pseudo-Invertible Decoder (APID)**




**APID**

- Residual normalizing flows + variational augmentations; fits the observational data
- Performs abduction–action–prediction differentiably and penalizes the curvature

← backup overview

# Backup | Epistemic uncertainty: efficiency & optimality

**Asymptotic normality**

- $\sqrt{n}\left(\hat{\psi}_n^{\text{A-IPTW}} - \psi\right) \rightsquigarrow \mathcal{N}(0, \sigma^2), \quad \sigma^2 = \text{Var}[\varphi_\psi(Z;\eta)] \quad \Rightarrow$
  $\boxed{\text{CI}_{1-\alpha} = \hat{\psi}_n^{\text{A-IPTW}} \pm z_{1-\alpha/2}\,\hat{\sigma}/\sqrt{n}}$
- Requires strong overlap and $\|\hat{\mu}_a - \mu_a\|\,\|1/\hat{\pi}_a - 1/\pi_a\| = o_{\mathbb{P}}(n^{-1/2})$ (cross-fitting)

**Local optimality**

- Semi-parametric efficiency bound:
  $\inf_{\delta>0} \liminf_{n\to\infty} \sup_{d_{\text{TV}}(\mathbb{P},\mathbb{Q})<\delta} n\,\mathbb{E}_{\mathbb{Q}^n}\left[(\hat{\psi}_n - \psi(\mathbb{Q}))^2\right] \geq \text{Var}[\varphi_\psi(Z;\eta)]$

**Global optimality**

- With black-box nuisance learners, first-order debiased estimators are **structure-agnostic minimax optimal** for APOs / ATE
- Higher-order corrections help only under extra structure (Balakrishnan et al., 2026; Jin & Syrgkanis, 2025)

**Finite-sample & Bayesian**

- Unknown propensity score: no uniformly valid finite-sample CIs without smoothness assumptions; known propensity + bounded outcome: Hoeffding-type CIs (conservative)
- Plug-in posteriors can fail the Bernstein–von Mises property unless the nuisances are smooth enough; remedies: prior / posterior corrections (Ray & van der Vaart, 2020; Yiu et al., 2025)

← backup overview

# Backup | Neyman-orthogonality & its consequences

**Two-stage learning**

- Stage 1: nuisance functions $\hat{\eta}$ (e.g., $\hat{\pi}_a, \hat{\xi}_a, \hat{\mathbb{F}}_a$), cross-fitted
- Stage 2: $\hat{g} = \arg\min_{g \in \mathcal{G}} \hat{\mathcal{L}}(g, \hat{\eta})$, oracle: $g^* = \arg\min_{g \in \mathcal{G}} \mathcal{L}(g, \eta)$

**Neyman-orthogonality**

- $\boxed{D_\eta D_g \, \mathcal{L}(g^*, \eta)[g - g^*, \hat{\eta} - \eta] = 0}$: the gradient of the risk at the oracle is first-order insensitive to nuisance errors (Foster & Syrgkanis, 2023)
- Construction: one-step bias correction of a plug-in risk with the EIF of the risk at fixed $g$

**Quasi-oracle efficiency**

- Under strong convexity, $\|\hat{g} - g^*\|_{\mathcal{G}}^2 \lesssim$ optimization error $+ \, \|\hat{\eta} - \eta\|_{\mathcal{H}}^4$ with a *second-order* nuisance term
- With nuisances at $o_{\mathbb{P}}(n^{-1/4})$, learning is asymptotically as good as with known nuisances (Nie & Wager, 2021)

**(Rate) double robustness**

- The second-order term is a *product*, e.g., $\|\hat{\xi}_a - \xi_a\|_{L_4}^2 \|\hat{\pi}_a - \pi_a\|_{L_4}^2$: a slowly converging nuisance is compensated by a fast one

**In this thesis**

- INFs: EIF of the moment condition • GDR-learners: orthogonal generative risks • AU-learner: EIF of Makarov bounds

← backup overview

# Backup | (Non-exhaustive) overview of approaches

| Source | Estimand / target | Identification | Estimation |
|---|---|---|---|
| **EU** | asymptotic variance of A-IPTW; APO/ATE posterior | point: POF | plug-in / robust variance estimators; prior & posterior corrections |
| **AU** | DPOs; CDPOs | point: POF | plug-in; semi-parametric density estimation; **INFs**; **GDR-learners** |
| **UU** | confounded APOs / ATE / CAPOs / CATE; representation-level CATE; ECOU/ECOT | partial: Manski bounds; sensitivity models (MSM etc.); **CSM** | plug-in (incl. **RICB bounds**); A-IPTW; TMLE; B-learner |
| **AU + UU** | CDF / quantiles / variance of the TE | partial: Makarov, Fréchet–Hoeffding bounds | plug-in; **AU-learner** |
| **EU + AU** | posterior predictive of POs; prediction intervals of POs | point: POF | amortized Bayesian inference; conformal prediction |
| **EU + UU** | confounded APO / ATE posterior | partial: Bayesian sensitivity models | plug-in posterior |
| **EU + AU + UU** | posterior predictive / prediction intervals of the TE; POs under confounding | partial: POF; MSM | amortized inference; conformal (sensitivity) prediction |

Adapted from Table 1 in the introduction of the dissertation (Melnychuk, 2026).

← backup overview

# Backup | INFs as a special case of GDR-learners

**GDR-learner with $V = \emptyset$**

- Choose an unconditional target model $g_a(y) = g(y; \beta_a)$ (no covariates, i.e., the DPO):

$$\hat{\mathcal{L}}_{\text{GDR}}(\beta_a) = \mathbb{P}_n \left\{ \frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a(X)} \log g(Y; \beta_a) + \left(1 - \frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a(X)}\right) \int \log g(y; \beta_a)\, \hat{\xi}_a(y \mid X)\, \mathrm{d}y \right\}$$

**INFs objective**

- The negative GDR risk equals the INFs (A-IPTW) loss:

$$-\hat{\mathcal{L}}_{\text{GDR}}(\beta_a) = \underbrace{\mathbb{E}_{\hat{Y} \sim \mathbb{P}_n\{\hat{\xi}_a(\cdot|X)\}} \big[ -\log g(\hat{Y}; \beta_a) \big]}_{\text{cross-entropy}} + \underbrace{\mathbb{P}_n \left\{ \frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a(X)} \Big( -\log g(Y; \beta_a) - \mathbb{E}_{Y \sim \hat{\xi}_a(\cdot|X)} \big[ -\log g(Y; \beta_a) \big] \Big) \right\}}_{\text{one-step bias correction}}$$

$\rightarrow$ maximizing the GDR risk = minimizing the INFs (A-IPTW) loss

**What GDR adds**

- Covariate-conditional targets $\mathbb{P}(Y[a] \mid V)$, $V \subseteq X$ (fairness / interpretability constraints)
- *Any* generative model (also implicit ones: GANs, VAEs, diffusion models) via sampling from $\hat{\xi}_a$
- Asymptotic optimality guarantees: quasi-oracle efficiency & rate double robustness

← backup overview

# Backup | INFs: design choices & efficiency

| | |
|---|---|
| **Why fully parametric?** | • Proper normalized density, direct sampling, and constant inference time<br>• The target flow does not average over the training data at test time (unlike semi-parametric plug-in estimators) |
| **Efficiency "wrt. the moment condition"** | • The density is infinite-dimensional (no EIF); we target the finite-dimensional projection parameter $\beta_a^*$, defined by $m(\beta_a) = 0$<br>• A-IPTW efficiency & double robustness hold for this moment condition (Kennedy et al., 2023), also for equivalence classes of deep model parameters |
| **Why normalizing flows?** | • Universal density approximators with tractable likelihood (neural spline flows: flexible & parsimonious)<br>• Conditional version via hypernetworks for the nuisance flow |
| **Stabilization** | • Noise regularization against overfitting of the conditional flow; two-step training with gradient blocking<br>• Exponential moving average (EMA) of target parameters for small mini-batches |
| **Integral in the loss** | • $\mathcal{L}_{\text{CCE}}(X; \beta_a) = -\int \log g(y; \beta_a)\, \hat{\mathbb{P}}(Y = y \mid X, A = a)\, \mathrm{d}y$ via quadrature ($d_y = 1$) or Monte Carlo ($d_y > 1$) |

← backup overview

# Backup | GDR-learners: are $o_{\mathbb{P}}(n^{-1/4})$ rates achievable?

| | |
|---|---|
| **Requirement** | • Quasi-oracle efficiency needs the nuisances (conditional outcome density $\xi_a$ and propensity score $\pi_a$) to converge at $o_{\mathbb{P}}(n^{-1/4})$, or their product at $o_{\mathbb{P}}(n^{-1/2})$ |
| **Classical rates** | • Conditional densities are functions $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$: non-parametric rates $O_{\mathbb{P}}\left(n^{-s/(2s+d_x+d_y)}\right)$, where $s$ is Hölder smoothness (Stone, 1982)<br>• Slow for large $d_x + d_y$ |
| **Deep learning** | • Deep networks can adapt to low-dimensional manifold structure and improve these rates (e.g., Schulte et al., 2025 for ATE nuisances; Melnychuk et al., 2026 for orthogonal learners)<br>• Convergence rates of deep conditional generative models are an active research area |
| **In practice** | • Even when rates are not attained, orthogonality reduces first-order sensitivity to nuisance errors<br>• Double robustness lets a well-estimated propensity score compensate a harder outcome model |

← backup overview

# Backup | AU-learner: sharpness, assumptions & low-sample regime

| | |
|---|---|
| **Sharpness** | • Makarov bounds are *pointwise* sharp: for each $\delta$, some coupling of $\mathbb{P}(Y[0] \mid x)$ and $\mathbb{P}(Y[1] \mid x)$ attains the bound<br>• The bounds themselves are CDFs, but not the CDF of a single coupling; analogous bounds exist for quantiles and for categorical / mixed outcomes |
| **Why not point identification?** | • Rank invariance / monotone BGMs identify the CDTE, but impose a common latent ranking across treatment arms (deterministic counterfactuals)<br>• Rarely credible and untestable from data |
| **Regularity** | • Pathwise differentiability needs finite argmax / argmin sets of the sup / inf convolutions; otherwise inference is non-regular (as for **the value of optimal policies**) |
| **Low-sample regime** | • With small $n$ or strong overlap violations (e.g., IHDP), RA- or plug-in learners can be competitive<br>• No data-driven way to pick the best low-sample learner without counterfactuals: the asymptotically optimal AU-learner is the pragmatic default; $\gamma$ ($0.25$ for CRPS) improves stability |

← backup overview

# Backup | AU-learner vs. conformal prediction & Bayesian methods

| | **AU-learner (Makarov bounds)** | **Conformal prediction for ITEs** (Lei & Candès, 2021; Alaa et al., 2023) |
|---|---|---|
| **Target** | bounds on the CDF / quantiles of $\mathbb{P}(Y[1] - Y[0] \mid x)$ | prediction interval $\hat{C}(X)$ with $\mathbb{P}\big(Y[1] - Y[0] \in \hat{C}(X)\big) \geq 1 - \alpha$ |
| **Uncertainty** | AU + UU, separated from EU | EU + AU + UU combined |
| **Guarantee** | asymptotic: orthogonality, quasi-oracle efficiency | finite-sample, marginal coverage (often conservative for ITEs) |
| **Output** | whole (bounded) distribution: probability of benefit, quantiles | one interval per coverage level |

- Bayesian / amortized approaches (e.g., causal prior-fitted networks) give posterior predictive distributions of the TE, but the UU part is driven by the prior over couplings

← backup overview

# Backup | Limitations of this thesis

| | |
|---|---|
| **Assumptions** | • Parts I & III rely on the standard POF assumptions (consistency, overlap, unconfoundedness)<br>• Part II relaxes identification in specific ways (representations, SCM curvature) |
| **Guarantees** | • Mostly asymptotic (efficiency, orthogonality); in low-sample regimes, simpler learners can be competitive |
| **Scope** | • EU of the learned distributions / bounds is not quantified (for that we would need to know how to quantify the EU of the target model)<br>• Static settings, binary (or categorical) treatments<br>• Makarov bounds for univariate outcomes; high-dimensional outcomes only for CDPOs |
| **Evaluation** | • Ground truth requires (semi-)synthetic benchmarks<br>• Model selection without counterfactuals remains an open research direction, especially for distributional and partially identified estimands |

← backup overview