Generalized Bayes for Causal Inference

Emil Javurek[*], **Dennis Frauen, Yuxin Wang, Stefan Feuerriegel**
LMU Munich
Munich Center of Machine Learning (MCML)

Abstract

Uncertainty quantification is central to many applications of causal machine learning, yet principled Bayesian inference for causal effects remains challenging. Standard Bayesian approaches typically require specifying a probabilistic model for the data-generating process, including high-dimensional nuisance components such as propensity scores and outcome regressions. Standard posteriors are thus vulnerable to strong modeling choices, including complex prior elicitation. In this paper, we propose a *generalized Bayesian framework for causal inference*. Our framework avoids explicit likelihood modeling; instead, we place priors directly on the causal estimands and update these using an identification-driven loss function, which yields *generalized posteriors* for causal effects. As a result, our framework turns existing loss-based causal estimators into estimators with full uncertainty quantification. Our framework is flexible and applicable to a broad range of causal estimands (e.g., ATE, CATE). Further, our framework can be applied on top of state-of-the-art causal machine learning pipelines (e.g., Neyman-orthogonal meta-learners). For Neyman-orthogonal losses, we show that the generalized posteriors converge to their oracle counterparts and remain robust to first-stage nuisance estimation error. With calibration, we thus obtain valid frequentist uncertainty even when nuisance estimators converge at slower-than-parametric rates. Empirically, we demonstrate that our proposed framework offers causal effect estimation with calibrated uncertainty across several causal inference settings. To the best of our knowledge, this is the first flexible framework for constructing generalized Bayesian posteriors for causal machine learning.

1 Introduction

Causal inference is widely used to assess treatment effects in order to compare interventions and inform decisions [9, 4, 18]. However, conclusions are typically not based on the point estimates of causal effects alone, but rather on considering uncertainty associated with estimating causal effects. For example, in medicine, a treatment that appears beneficial on average may still be clinically unacceptable if the uncertainty is large enough to imply a substantial probability of harm [48, 22]. As a result, many applications of causal inference, require not only estimates of causal effects but also rigorous *uncertainty quantification*.

Bayesian inference provides a natural framework for uncertainty quantification by computing a posterior distribution over the quantity of interest. However, in causal inference, standard Bayesian approaches are subject to fundamental *challenges*: existing approaches typically require specifying a probabilistic model of the observational data-generating process, including nuisance components such as propensity scores and outcome regressions (see Section 2 for a detailed overview). This is cumbersome and the resulting posterior is fragile to the modeling choices. Moreover, a prior must be placed on the nuisance components, which can interact with the likelihood in unintended ways, such as introducing regularization-induced confounding [17].

[*]Correspondence to: emil.javurek@lmu.de

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

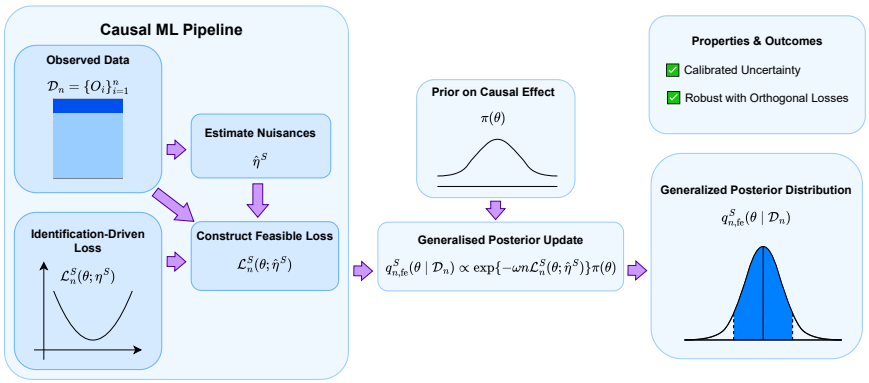

Figure 1: **Overview of the pipeline for generalized Bayesian inference:** In the standard causal ML pipeline, one observes the data $\mathcal{D}_n$, chooses an identification-driven loss, estimates the corresponding nuisances, and thus constructs a fitting objective. Here, we further elicit a prior on the causal effect and together construct a generalized updating rule yielding the generalized posterior distribution for the causal effect of interest. This approach can be calibrated to yield frequentist-valid uncertainty and be made robust to nuisance estimation error via Neyman-orthogonal losses.

In this paper, we propose a ***generalized Bayesian framework for causal inference*** that avoids explicit likelihood modeling.[2] Conceptually, we follow recent works on generalized Bayes [25] and view Bayesian inference *not* just through the Bayes rule formula, but frame it as a principled belief-updating procedure that naturally extends to loss-based formulations more compatible with modern machine learning (ML). In this framework, we thus avoid specifying a probabilistic model of the full observational data-generating process; instead, we (i) place priors directly on causal estimands and (ii) update these using identification-driven loss functions for which the minimizers recover the targeted causal effect. This construction yields generalized posteriors for causal effects, without requiring priors or likelihoods for high-dimensional nuisance components. This addresses key challenges of likelihood-based Bayesian approaches, which can be sensitive to nuisance component modeling misspecification. Consequently, our approach brings the following benefits:

• **Flexibility.** Our framework is applicable to a wide range of causal estimands. Throughout our paper, we illustrate our framework using two canonical estimands from the causal inference literature: (i) the average treatment effect (ATE) and (ii) the conditional average treatment effect (CATE) settings. The proposed framework extends naturally to any other identified causal estimand that can be estimated with a loss. Moreover, our framework can be applied on top of state-of-the-art causal ML pipelines, in particular Neyman-orthogonal meta-learners such as the DR-learner [20, 44]. Altogether, our framework provides a general recipe for transforming existing loss-based causal estimators into estimators with full uncertainty quantification.

• **Theoretical guarantees.** For Neyman-orthogonal losses, we show that the generalized posteriors have several favorable properties. In particular, the generalized posterior converges to the oracle counterpart and remains robust to first-stage nuisance estimation error, even when nuisance estimators converge at slower-than-parametric rates (analogous to state-of-the-art Neyman-orthogonal meta-learners for point estimates). It follows that the estimated generalized posterior also satisfies a Bernstein-von Mises limit, yielding asymptotically valid frequentist uncertainty quantification. This establishes how orthogonal statistical learning applies to generalized Bayesian causal inference.

• **Practical advantages.** *First*, our framework enables the placement of priors directly on the causal estimand of interest, which allows prior information to be specified transparently and avoids the need to encode prior beliefs indirectly through high-dimensional nuisance models, where their effect on the causal estimand is difficult to interpret or control. *Second*, this loss-based formulation makes it possible to characterize and control how nuisance estimation error propagates into posterior uncertainty through properties of the loss such as Neyman-orthogonality. *Third*, the resulting generalized posteriors can be calibrated to match frequentist uncertainty and thus achieve coverage.

Our main **contributions** are:[3] **(1)** We introduce a generalized Bayesian framework for causal inference, which is applicable to state-of-the-art causal ML pipelines. **(2)** We prove that, for Neyman-

[2] By "generalized", we mean Bayesian updating based on a loss function rather than a likelihood, following the generalized (specifically, Gibbs) posterior framework of Bissiri et al. [5]. Standard Bayesian inference is recovered as a special case when the loss is the negative log-likelihood.

[3] Code is available at `https://github.com/EmilJavurek/Generalised-Bayes-for-Causal-Inference`

orthogonal losses, the resulting generalized posterior is robust to first-stage nuisance estimation error and thus offers explicit convergence guarantees for the posterior. **(3)** We empirically demonstrate that our framework yields frequentist-calibrated uncertainty quantification across several causal inference settings. Importantly, the goal of our framework is *not* to approximate a standard likelihood-based Bayesian posterior, but to do loss-based updates of prior beliefs that yield *generalized* posteriors with valid frequentist uncertainty. To the best of our knowledge, ours is the first framework that constructs generalized posteriors for causal inference using identification-driven losses with formal robustness guarantees wrt. the nuisance estimation error.

2 Related Work

Causal ML: Recent work combines causal inference with machine learning to develop flexible methods for estimating a wide range of causal estimands, including the ATE [8, 21, 45] and CATE [10, 30, 20]. State-of-the-art methods are typically *Neyman-orthogonal meta-learners* (e.g., the DR-learner [43, 20], R-learner [30]), which use flexible machine learning models to estimate nuisance components and then target the causal effect in the second stage with a Neyman-orthogonal loss that ensures first-order insensitivity to nuisance estimation error [10, 8, 20].

Generalized Bayesian inference: Generalized Bayesian inference constructs posterior distributions by updating a prior using a loss function, rather than a likelihood. Conceptually, we may understand this generalization through the role the likelihood/losses play in the belief updating procedure: *both simply quantify the discrepancy between a candidate parameter value and the observed data*. Bissiri et al. [5] showed that this loss-based update yields a principled posterior under a decision-theoretic formulation, where standard Bayesian inference is recovered as a special case when the loss is the negative log-likelihood. Further works have extended the theory of generalized Bayes [27, 23, 24, 13]; *however, to the best of our knowledge, no existing work has explored the extensions of generalized Bayes theory to causal inference - specifically, the question of employing losses depending on estimated nuisances.*

Likelihood-based Bayesian causal inference:

A large body of Bayesian causal inference proceeds by specifying a probabilistic model for (all parts of) the observed data distribution and performing posterior inference under this model. Many approaches leverage flexible Bayesian nonparametric priors over high-dimensional nuisance components [e.g., 2, 33, 29]; see Linero and Antonelli [26] for an overview. *However, this entails modeling high-dimensional nuisance components and specifying corresponding priors, introducing challenges for robustness and prior elicitation (which we discuss further in Section 4).*

In parallel, Bayesian approaches have been developed for specific settings but tied to specific model classes [e.g., 49, 46, 16, 7, 49, 1]. These methods typically embed causal structure into a chosen probabilistic model (e.g., multi-task Gaussian process models for CATE estimation, which impose Gaussian process priors; Alaa and van der Schaar [1]). *However, such approaches are inherently model-specific and their usefulness hinges on the model class (the inductive bias) matching the problem at hand.*

A key challenge for Bayesian causal inference is achieving robustness to nuisance estimation error. In joint Bayesian models, posterior feedback between outcome and propensity-score components can undermine robustness under misspecification, motivating modular and cut-posterior approaches [41, 36, 19], including their extension to generalized Bayesian inference [11]. Other approaches pursue reliable uncertainty quantification for causal functionals through tailored doubly robust constructions or semiparametric posterior corrections [6, 32, 40, 3, 47]. *However, these approaches do not provide a general construction applicable across varying causal settings or estimands.*

We elaborate on the difficulties of standard likelihood-based Bayesian approaches in Section 4. Extended related work can be found in Appendix C.

Research gap: Overall, existing Bayesian causal methods largely rely on model- or estimand-specific constructions and often lack robustness to nuisance misspecification, which limits their applicability as a general framework for uncertainty quantification. A unified framework across estimands and models for (generalized) Bayesian causal inference is still missing.

3 Problem Setup

Notation: We denote random variables by capital letters X, A, Y, Z, O and their realizations by small letters x, a, y, z, o from domains $\mathcal{X}, \mathcal{A}, \mathcal{Y}, \mathcal{Z}, \mathcal{O}$. Let $\mathbb{P}(O)$ denote a distribution of some random variable O, and let $p(O = o)$ be a corresponding density or probability mass function. Let $\mathcal{P}(\mathcal{O})$ denote the set of all probability distributions over $\mathcal{O}$. We write $\mathbb{E}_{\mathbb{P}}$ to denote expectations with respect to the distribution $\mathbb{P}$ and $\mathbb{E}_n$ to denote the corresponding empirical mean over an observed sample of size n from $\mathbb{P}$. We use the potential outcomes framework [35].

Setting: We consider a *causal inference setting* given by a tuple $\mathcal{C} = \big(O, \mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}, Q\big)$, where O collects observed variables containing at least the treatment A and the outcome Y; $(\mathbb{P}_{\text{obs}}, \mathbb{P}_{\text{int}}) \in \mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$ are paired observational/interventional distributions over O corresponding to an intervention on A; and $Q(\mathbb{P}_{\text{int}}) \in \Theta$ is the causal estimand of interest.

The **causal estimand** $Q(\mathbb{P}_{\text{int}})$ is said to be *identifiable* if, under identification assumptions specific to the setting $\mathcal{C}$, there exists a measurable functional $\bar{Q}$ such that $Q(\mathbb{P}_{\text{int}}) = \bar{Q}(\mathbb{P}_{\text{obs}}) \equiv \theta$. Given an observed dataset $\mathcal{D}_n = \{O_i\}_{i=1}^n \sim \mathbb{P}_{\text{obs}}$, we estimate the identified causal estimand via $\theta \approx \hat{\theta}(\mathcal{D}_n)$.

Estimation: Given an observed dataset $\mathcal{D}_n = \{O_i\}_{i=1}^n \sim \mathbb{P}_{\text{obs}}$ from setting $\mathcal{C}$, we aim to estimate the identified causal estimand θ via an estimator $\hat{\theta}(\mathcal{D}_n)$. For a given estimand θ, there are typically multiple valid estimation strategies; here, we focus on a broad class of M-estimation methods. Formally, a strategy $S \in \mathcal{S}$ obtains a point-estimator $\hat{\theta}^S$ by a loss minimization problem

$$\hat{\theta}^S \in \arg\min_{\theta \in \Theta} \mathcal{L}_n^S(\theta) \equiv \arg\min_{\theta \in \Theta} \mathbb{E}_n\big[\ell^S(O; \theta)\big], \tag{1}$$

where $\mathcal{L}_n^S(\theta)$ is the empirical loss (using $\mathcal{D}_n$), and $\ell^S(O; \theta)$ is the per-observation loss.

Nuisances: In causal inference, estimation strategies S often depend on additional *nuisances* $\eta^S = \eta^S(\mathbb{P}_{\text{obs}})$. These nuisances (e.g., propensity scores, outcome regressions or conditional densities) are not themselves the target causal estimand, but are used as part of the estimation procedure. Accordingly, we define an estimation strategy S to include both the nuisances η^S and an estimator $\hat{\eta}^S = \hat{\eta}^S(\mathcal{D}_n)$ for the nuisances. Correspondingly, we write $\ell^S(O; \theta, \hat{\eta}^S)$ for the nuisance dependent per-observation loss. At the population level, losses are typically motivated by an *oracle* score or pseudo-outcome constructed using the true nuisance η^S. In practice, replacing η^S by $\hat{\eta}^S$ is unavoidable and induces *nuisance estimation error*, which can dominate the error of $\hat{\theta}^S$ in complex/high-dimensional settings. Our framework is agnostic to how nuisance estimators are obtained, nevertheless, we discuss options in Appendix B.2. A central goal of our later theory will be to characterize how such nuisance error propagates downstream into the generalized posterior.

Our framework applies to a broad class of causal estimands that fall under the above causal inference setting; to make the exposition concrete, we introduce a set of running examples below, which we use throughout the paper to illustrate the construction, derivations, and theoretical results.

■ **Example 1 (back-door ATE).** Consider $O = (X, A, Y) \sim \mathbb{P}_{\text{obs}}$ where X are covariates, $A \in \{0, 1\}$ is a binary treatment, and Y is an outcome. Let $Y(a)$ denote the potential outcome under treatment $A = a$ and write $\mathbb{P}_{\text{int}}$ for the interventional distribution induced by setting $A \leftarrow a$. The *average treatment effect* is given by $\theta_{\text{ATE}} = \mathbb{E}\big[Y(1) - Y(0)\big]$. We impose standard identification assumptions [34]: (i) *consistency* $Y = Y(A)$, (ii) *positivity* $\mathbb{P}_{\text{obs}}(A = a \mid X = x) > 0$ a.s. for $a \in \{0, 1\}$, and (iii) *unconfoundedness* $(Y(0), Y(1)) \perp\!\!\!\perp A \mid X$. Under these conditions, the target is identified as functional of $\mathbb{P}_{\text{obs}}$. Writing $m_a(x) = \mathbb{E}_{\mathbb{P}_{\text{obs}}}[Y \mid A = a, X = x]$, we have $\theta_{\text{ATE}} = \mathbb{E}_{\mathbb{P}_{\text{obs}}}\big[m_1(X) - m_0(X)\big]$. Scalar estimands such as the ATE often admit closed-form minimizers, such as $\hat{\theta}^S = \frac{1}{n}\sum_{i=1}^n \hat{Y}^S(O_i)$ with strategy $S \in \{\text{RA}, \text{IPW}, \text{AIPW}^4\}$ determining the pseudo-outcome form $\hat{Y}^S$. With finite-dimensional problems where the estimator admits a closed-form expression, the estimation strategy can nevertheless be represented as the minimizer of an appropriate loss, provided the loss is from a Fisher-consistent family of losses (e.g., squared loss for mean-type estimands, absolute losses for medians, or pinball losses for quantiles) [15]. Since the ATE is a mean, we may thus define a loss $\ell^S(O_i; \theta, \hat{\eta}^S) = (\hat{Y}^S(O_i; \hat{\eta}^S) - \theta)^2$ for any of the above strategies S.

[4] For the ATE, the pseudo-outcomes $\hat{Y}^{\text{AIPW}}$ from the AIPW strategy directly correspond to the $\hat{Y}^{\text{DR}}$ from CATE estimation.

■ **Example 2 (back-door CATE).** In the same setting as Example 1, we consider the *conditional average treatment effect* (CATE), defined as $\theta_{\mathrm{CATE}}(x) = \mathbb{E}\big[Y(1) - Y(0) \mid X = x\big]$. Under the same assumptions, the CATE is identified by the observational conditional mean functions as $\theta_{\mathrm{CATE}}(x) = m_1(x) - m_0(x)$. A common approach for CATE estimation is to construct a pseudo-outcome $\hat{Y}^S(O;\eta^S)$ satisfying $\mathbb{E}_{\mathbb{P}_{\mathrm{obs}}}[\hat{Y}^S(O;\eta^S) \mid X = x] = \theta_{\mathrm{CATE}}(x)$ and fit $\hat{\theta}(\cdot)$ by regression using the squared error loss $\ell^S(O;\theta,\hat{\eta}^S) = \big(\hat{Y}^S(O;\hat{\eta}^S) - \theta(X)\big)^2$. Typical choices for strategy include regression adjustment (RA), $\hat{Y}^{\mathrm{RA}} = \hat{m}_1(X) - \hat{m}_0(X)$; inverse propensity weighting (IPW), $\hat{Y}^{\mathrm{IPW}} = \frac{AY}{\hat{e}(X)} - \frac{(1-A)Y}{1-\hat{e}(X)}$; and doubly robust estimation (DR), $\hat{Y}^{\mathrm{DR}} = \hat{Y}^{\mathrm{RA}} + \big(\frac{A}{\hat{e}(X)} - \frac{1-A}{1-\hat{e}(X)}\big)(Y - \hat{m}_A(X))$.

4 Why standard Bayesian approaches for causal inference are difficult

The naïve approach to Bayesian inference of a causal query θ (in a fixed causal setting $\mathcal{C}$) would be to simply follow the classical Bayes rule $q(\theta|\mathcal{D}_n) \propto \mathbb{P}_{\mathrm{obs}}(\mathcal{D}_n|\theta)\pi(\theta)$[5]. While specifying a prior $\pi(\theta)$ directly on $\theta \in \Theta$ does not pose any issues – and is in fact desirable for encoding belief into prior – the likelihood $\mathbb{P}_{\mathrm{obs}}(\mathcal{D}_n|\theta)$ is problematic. In causal settings, θ typically does not fully determine the observed data distribution $\mathbb{P}_{\mathrm{obs}}$, so $\mathbb{P}_{\mathrm{obs}}(\mathcal{D}_n|\theta)$ is not directly defined without further modeling of the data-generating process. For example:

■ **Example 1 (back-door ATE).** Many different distributions $\mathbb{P}_{\mathrm{obs}}$ of (X, A, Y) induce the *same* ATE θ_{ATE}. As a result, θ_{ATE} does *not* uniquely determine a value of the likelihood for (X, A, Y), i.e. $\mathbb{P}_{\mathrm{obs}}(\mathcal{D}_n \mid \theta_{\mathrm{ATE}})$ is *not* well-defined.

Instead, a proper likelihood is defined as $p(\mathcal{D}_n \mid \mathbb{P}_{\mathrm{obs}})$, with a corresponding prior $\pi(\mathbb{P}_{\mathrm{obs}})$ over the space of data-generating distributions. This induces the posterior distribution

$$q(\theta \mid \mathcal{D}_n) = \int \delta_\theta(\bar{Q}(\mathbb{P}_{\mathrm{obs}}))p(\mathbb{P}_{\mathrm{obs}} \mid \mathcal{D}_n)\, \mathrm{d}\mathbb{P}_{\mathrm{obs}}, \tag{2}$$

which is a marginalization of the joint posterior $p(\mathbb{P}_{\mathrm{obs}}|\mathcal{D}_n) \propto \pi(\mathbb{P}_{\mathrm{obs}})p(\mathcal{D}|\mathbb{P}_{\mathrm{obs}})$, and where $\delta_\theta(\bar{Q}(\mathbb{P}_{\mathrm{obs}}))$ denotes a Dirac delta of $\bar{Q}(\mathbb{P}_{\mathrm{obs}})$ at θ.

To make the prior distribution and likelihood at the level of $\mathbb{P}_{\mathrm{obs}}$ tractable, we restrict the space of possible data-generating distributions to a (semi-) parameterized subset $\mathcal{Q}_{\mathrm{obs}} \subset \mathcal{P}_{\mathrm{obs}}$. For example:

■ **Examples 1 & 2 (back-door ATE & CATE):** In the back-door adjustment settings, we may parameterize a model of the data-generating process $\mathbb{P}_{\mathrm{obs}}$ by $\mathbb{P}_{\mathrm{obs},\xi}(O) = p_{\xi_x}(X)p_{\xi_a}(A \mid X)p_{\xi_y}(Y \mid A, X)$ from a corresponding model class $Q_{\mathrm{obs},\xi} = \big\{\mathbb{P}_{\mathrm{obs},\xi}(O) \mid \xi = (\xi_x, \xi_a, \xi_y) \in \Xi\big\}$, together with some prior $\pi(\xi) \in \mathcal{P}(\Xi)$ on the parameters ξ. Notice that ξ_a and ξ_y actually parametrize the propensity and outcome regression nuisances. The standard Bayes approach thus forces us to specify a prior over nuisances and to compute a posterior for the entire ξ, which will typically be higher-dimensional than θ, possibly even infinite-dimensional. The posterior for θ must then finally be computed as a marginalization of the overall posterior over all (model) parameters ξ of the data-generating process.

As illustrated by the above example, several practical difficulties arise even in seemingly simple causal settings. In particular, the following issues arise:

(1) Model misspecification: The restricted model class $\mathcal{Q}_{\mathrm{obs}} \subset \mathcal{P}_{\mathrm{obs}}$ may fail to contain the true observational distribution $\mathbb{P}^\star_{\mathrm{obs}}$, especially if the model class is overly restrictive or otherwise misspecified. In this case, the likelihood is misspecified and the correct Bayesian posterior is *not* obtained. Nevertheless, such restrictions are often unavoidable in practice for computational and regularization reasons.

(2) Priors on high-dimensional nuisance functions. As illustrated in the back-door example, specifying a likelihood requires placing priors on nuisance components such as propensity scores and outcome regressions. Even after restricting $\mathcal{Q}_{\mathrm{obs}}$, these nuisance objects are typically high-dimensional functions, for which posterior convergence and Bernstein–von Mises approximation are *not* guaranteed *unless strong additional assumptions are enforced* [12].

(3) Sensitivity to nuisance priors, including feedback problems. The (marginalized) posterior over the causal estimand θ depends on an appropriate prior specification for the nuisance components but

[5] By $\propto$ we denote proportionality up to a multiplicative constant.

correct prior elicitation is difficult. For example, even when the propensity and outcome nuisances are given independent priors, a joint likelihood couples their posteriors; outcome information can then feed back into propensity estimation, which may harm balance and robustness under misspecification (the feedback problem) [41, 32].

(4) Lack of direct control over the induced prior on the causal estimand. Placing a prior on $\mathbb{P}_{\text{obs}}$ induces also a prior on the causal estimand θ via a non-trivial mapping $\pi(\theta) = \int \delta_{\bar{Q}(\mathbb{P}_{\text{obs}})}(\theta)\pi(\mathbb{P}_{\text{obs}})\ \mathrm{d}\mathbb{P}_{\text{obs}}$. As a consequence, it is difficult – if not impossible – to directly elicit prior beliefs about θ. This is critical if we want to follow the philosophy of Bayesian updating of beliefs; following the standard Bayes rule formula in causal inference settings may thus make the task of eliciting a prior belief on θ difficult.

With the understanding of how standard Bayesian inference in causal inference is made difficult by nuisance functions, we now turn to generalized Bayes for an alternative solution.

5 Method (Generalized Bayes)

We proceed in four steps. In §5.1, we introduce generalized posteriors based on identification-driven losses, which define a principled Bayesian update of belief for causal estimands. In §5.2, we turn to the realistic case where nuisances must be estimated and, in §5.3, we present an algorithm for computing the corresponding feasible generalized posteriors. In §5.4, we provide our main theoretical result: for Neyman-orthogonal losses with cross-fitting, the feasible posterior converges to its oracle counterpart and yields valid uncertainty quantification.

5.1 Generalized posteriors

To define a generalized posterior, we consider a fixed causal setting $\mathcal{C}$ with an identified estimand $\theta^\star \in \Theta$ and an estimation strategy S with loss $\mathcal{L}_n^S(\theta; \eta^S)$. Let $\pi(\theta) \in \mathcal{P}(\Theta)$ denote a prior distribution on the causal estimand, and let $\omega > 0$ be a calibration parameter. Then, the corresponding *generalized posterior* [5] is defined[6] as

$$q_n^S(\theta \mid \mathcal{D}_n) \propto \exp\{-\omega n \mathcal{L}_n^S(\theta; \eta^S)\}\pi(\theta). \tag{3}$$

This construction turns any loss-based estimator of a causal estimand $\theta^\star$ into a posterior distribution over Θ *without* specifying a likelihood or modeling the observational distribution $\mathbb{P}_{\text{obs}}$ (observe that $\mathcal{L}$ set to $\frac{1}{\omega}$-scaled negative log-likelihood recovers standard Bayes posterior; for detailed interpretation see Appendix B.1). Equivalently, the same generalized posterior can be characterized as the solution to an optimization problem over distributions

$$q_n^S(\theta \mid \mathcal{D}_n) = \operatorname*{arg\,min}_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim q}\left[\omega n \mathcal{L}_n^S(\theta; \eta^S)\right] + \mathrm{KL}(q \parallel \pi) \right\}$$

where $\mathrm{KL}(\cdot \parallel \cdot)$ denotes the Kullback–Leibler divergence [25]. This offers a variational formulation, which is particularly useful in practice, as simply replacing $\mathcal{P}(\Theta)$ by a desired variational family yields a tractable variational objective for approximating the generalized posterior.

5.2 Feasible posteriors

The generalized posterior defined above assumes that the nuisance components entering the loss are known. In practice, however, these nuisances must be estimated from data. Let $\mathbb{P}_{\text{obs}}$ denote the true observational distribution, and write $\eta_0^S := \eta^S(\mathbb{P}_{\text{obs}})$ for the corresponding true (oracle) nuisances of the estimation strategy S. We define the *oracle* generalized posterior as $q_{n,\text{or}}^S(\theta \mid \mathcal{D}_n) := q_n^S(\theta \mid \mathcal{D}_n; \eta_0^S)$. In practice, η_0^S is unknown and must be estimated, which yields the *feasible* generalized posterior $q_{n,\text{fe}}^S(\theta \mid \mathcal{D}_n) := q_n^S(\theta \mid \mathcal{D}_n; \hat{\eta}^S)$.

5.3 Algorithm

To compute the feasible generalized posterior in practice, we propose a generic algorithmic procedure (see Algorithm 1). The algorithm is agnostic to the specific causal estimand (e.g., ATE, CATE) and estimation strategy (e.g., Neyman-orthogonal estimators). The algorithm employs cross-fitting to estimate nuisance components on held-out data, which mitigates overfitting and is needed for Neyman-orthogonality later.

[6] This version of generalized posteriors is also sometimes referred to as Gibbs posteriors. For other generalization of the Bayes posterior, see, for example, [25].

Choosing the calibration parameter ω: The calibration parameter ω is not specified by either the prior or the loss, and must be chosen separately. To achieve frequentist-valid uncertainty quantification, we follow standard practice in generalized Bayes and tune ω, so that a chosen $(1-\alpha)$ posterior credible set has approximate $(1-\alpha)$ repeated-sampling coverage. Concretely, we use the bootstrap calibration scheme of Syring and Martin [42]: for a candidate ω, we construct the credible region $C_{\omega,\alpha}(\mathcal{D}_n)$, estimate its coverage by nonparametric bootstrap resampling (by checking containment of the point estimate $\hat{\theta} = \arg\min_{\theta\in\Theta} \mathcal{L}_n^S(\theta, \hat{\eta}^S)$), and we update ω via one-dimensional search or stochastic approximation until the empirical coverage matches $1-\alpha$. We refer to Syring and Martin [42] for details.

Algorithm 1: Generalized posteriors for causal inference

Input: Data $\mathcal{D}_n = \{O_i\}_{i=1}^n$; estimation strategy S (loss ℓ^S and nuisance learner); prior $\pi(\theta)$; number of folds K; calibration rule $\mathsf{Calibrate}_\omega$; posterior engine (i.e., closed-form, variational inference, or MCMC sampling).
Output: $q_{n,\mathrm{fe}}^S$

1: Split indices $\{1,\dots,n\}$ into folds $\{I_k\}_{k=1}^K$.
2: **for** $k = 1,\dots,K$ **do**
3: Fit nuisances $\hat{\eta}^{S,(-k)}$ on $\{O_i : i \notin I_k\}$.
4: **end for**
5: Form the cross-fitted empirical loss

$$\mathcal{L}_n^S(\theta; \hat{\eta}^S) := \frac{1}{n}\sum_{k=1}^K \sum_{i\in I_k} \ell^S\big(O_i; \theta, \hat{\eta}^{S,(-k)}\big).$$

6: Calibrate the parameter $\omega \leftarrow \mathsf{Calibrate}_\omega(\mathcal{D}_n, \mathcal{L}_n^S, \pi)$
7: Construct the generalized variational objective

$$\mathcal{J}_{n,\mathrm{fe}}^S(q) := \omega n\, \mathbb{E}_{\theta\sim q}\big[\mathcal{L}_n^S(\theta; \hat{\eta}^S)\big] + \mathrm{KL}(q \parallel \pi).$$

8: Compute $q_{n,\mathrm{fe}}^S \in \arg\min_{q\in\mathcal{Q}} \mathcal{J}_{n,\mathrm{fe}}^S(q)$ using the posterior engine
9: **return** $q_{n,\mathrm{fe}}^S$ and posterior summaries (mean, credible interval).

When the object of interest is infinite-dimensional (e.g., a CATE function $\theta(\cdot)$), Bernstein–von Mises results and calibration guarantees only typically apply to finite-dimensional functionals $F(\theta)$ (such as finite set of point evaluations, the mean, the median, etc.). In these cases, we calibrate ω for the reported functional $F(\theta)$ using the same procedure. For squared-error losses, the generalized posterior corresponds to a Gaussian working likelihood with variance ω^{-1}. Hence, a convenient plug-in choice is therefore $\omega = \widehat{\mathrm{Var}}(\hat{Y}^S)^{-1}$, where $\hat{Y}^S$ denotes the pseudo-outcome.

The choices of prior $\pi(\theta)$ and number of cross-fitting folds K are discussed in Appendix B.3 and Appendix B.4, respectively.

5.4 Theoretical results

Our central methodological question is: *When does the feasible generalized posterior $q_{n,\mathrm{fe}}^S$ behave like the oracle generalized posterior $q_{n,\mathrm{or}}^S$ despite the nuisance estimation?*

The answer depends on how the nuisance estimation error enters the loss $\mathcal{L}_n^S(\theta; \hat{\eta}^S)$. State-of-the-art methods for causal ML [e.g., 8, 30, 20] address this challenge by constructing *Neyman-orthogonal* losses [10]. For such orthogonal losses, the score $D_\theta \mathcal{L}_n^S$ – i.e., which determines the first-order optimality condition and the corresponding gradient updates – is insensitive to local perturbations in the nuisances D_{η^S} around the true value η_0^S, i.e.,

$$D_{\eta^S} D_\theta \mathcal{L}_n^S(\theta^\star; \eta_0^S)[\theta - \theta^\star, \eta - \eta_0^S] = 0. \tag{4}$$

Neyman-orthogonality makes the leading bias term $\hat{\eta}^S - \eta_0^S$ vanish, so that the estimation of θ can remain $\sqrt{n}$-consistent and asymptotically normal, even when the nuisances are learned at slower, nonparametric rates (up to a product-rate condition; such as those of flexible machine learning models) [10].

In Theorem 1, we show that leveraging the same orthogonality principle for constructing generalized posteriors yields an analogous result: the feasible posterior converges to the oracle even at less-than-parametric rates of convergence of the nuisances.

Theorem 1 (Posterior stability under orthogonal losses (informal))**.** *Under the regularity conditions detailed in Appendix A, assume that: (i) the loss is Neyman orthogonal at $(\theta^\star, \eta_0^S)$ in the sense of Eq. 4; and (ii) $\hat{\eta}^S$ is obtained by sample splitting/cross-fitting and satisfies $\|\hat{\eta}^S - \eta_0^S\| =: r_n \to 0$. Then, the feasible generalized posterior is asymptotically close to the oracle one at the rate*

$$\mathrm{TV}\big(q_{n,\mathrm{fe}}^S,\ q_{n,\mathrm{or}}^S\big) \ =\ O_{\mathbb{P}}\big(\sqrt{n}\, r_n^2\big) \ +\ o_{\mathbb{P}}(1), \tag{5}$$

where TV *denotes total variation on the target of inference: either (a) the full parameter $\theta \in \mathbb{R}^d$ (finite-dimensional case), or (b) the induced posterior of any fixed finite-dimensional projection $T(\theta) \in \mathbb{R}^m$ when θ is infinite-dimensional (e.g. pointwise CATE values at fixed covariate points).*

In particular, if $r_n = o(n^{-1/4})$, then TV $\to 0$ *in probability and the feasible posterior inherits the corresponding oracle BvM limit.*
Proof. See Appendix A for formal statement and proof. □

Practical implications of the assumptions. The assumptions underlying Theorem 1 draw on established conditions from orthogonal statistical learning [10] and generalized-posterior Bernstein–von Mises theory [28], combining nuisance estimation requirements with regularity and posterior concentration conditions. In practice, these conditions require sufficiently accurate nuisance estimation and well-behaved losses with stable local curvature; when they fail, the posterior-stability and associated uncertainty guarantees need not hold.

The rate condition $r_n = o(n^{-1/4})$ matches the familiar threshold from orthogonal learning used in double ML theory [8]: the nuisance estimators may converge substantially slower than $n^{-1/2}$, while still yielding asymptotically valid uncertainty for low-dimensional targets. We have thus extended the theoretical guarantees of orthogonal statistical learning [10] from point estimation to their corresponding generalized posteriors.

Remark 1 (Why this fails without orthogonality)**.** *For non-orthogonal losses, the analogous* TV *of Eq.* (5) *typically scales as $O_{\mathbb{P}}(\sqrt{n}\, r_n)$ rather than $O_{\mathbb{P}}(\sqrt{n}\, r_n^2)$, which requires nuisance estimators to converge at near-parametric rates in order to keep posterior perturbations small. This explains why generalized posteriors for causal inference built from naïve plug-in losses can be unstable in modern settings with high-dimensional and flexible ML-based nuisance estimators.*

■ **Example 1 (back-door ATE).** We illustrate the robustness of generalized posteriors by deriving an analytic solution for the back-door adjusted ATE setting using the squared-error loss constructed from the AIPW pseudo-outcome $\hat{Y}^{\mathrm{AIPW}}$. Specifically, let us consider the loss

$$\mathcal{L}_n^{\mathrm{AIPW}}(\theta; \hat{\eta}) = \tfrac{1}{2n} \sum_{i=1}^{n} \big(\hat{Y}^{\mathrm{AIPW}}(O_i; \hat{\eta}) - \theta\big)^2. \tag{6}$$

The corresponding loss minimizer is the usual AIPW estimator, i.e., $\hat{\theta}^{\mathrm{AIPW}} = \arg\min_\theta \mathcal{L}_n^{\mathrm{AIPW}}(\theta; \hat{\eta}) = \frac{1}{n}\sum_{i=1}^n \hat{Y}_i^{\mathrm{AIPW}}$. The oracle pseudo-outcome $\hat{Y}^{\mathrm{AIPW}}(\cdot, \eta)$ coincides with the efficient influence function of θ^{AIPW} of the ATE, implying that the estimator $\hat{\theta}^{\mathrm{AIPW}}$ is asymptotically unbiased and statistically efficient.

The corresponding feasible generalized posterior is given by

$$q_{n,\mathrm{fe}}^{\mathrm{AIPW}}(\theta \mid \mathcal{D}_n) \propto \exp\Big\{-\omega n\, \mathcal{L}_n^{\mathrm{AIPW}}(\theta; \hat{\eta})\Big\}\, \pi(\theta) = \exp\left\{-\tfrac{\omega}{2} \sum_{i=1}^{n} (\hat{Y}_i^{\mathrm{AIPW}} - \theta)^2\right\} \pi(\theta), \tag{7}$$

which allows for a closed-form solution under a normal prior $\pi(\theta) = \mathcal{N}(m_0, s_0^2)$. In that case, the posterior is exactly normal, i.e.,

$$q_{n,\mathrm{fe}}^{\mathrm{AIPW}}(\cdot \mid \mathcal{D}_n) = \mathcal{N}(m_p, s_p^2), \text{ where } \quad s_p^2 = \big(s_0^{-2} + \omega n\big)^{-1},\ m_p = s_p^2\big(s_0^{-2} m_0 + \omega n\, \hat{\theta}^{\mathrm{AIPW}}\big).$$

In particular, under a diffuse prior $s_0^2 \to \infty$, we have $m_p \to \hat{\theta}^{\mathrm{AIPW}}$ and $s_p^2 \to (\omega n)^{-1}$, meaning the generalized posterior centers exactly at the classical orthogonal point estimate (while ω controls its spread).

Importantly, the only data-dependent term of the analytical solution of the posterior parameters m_p, s_p^2 is the point-estimate $\hat{\theta}^{\mathrm{AIPW}}$, i.e., the minimizer of $\mathcal{L}_n^{\mathrm{AIPW}}(\theta; \hat{\eta})$. As such, we can see how Neyman-orthogonality of the loss directly translates into convergence guarantees for the generalized posterior under nuisance estimation error. Our theorem shows this generally, even when no analytical solution to the posterior is available.

Finally, note that the analytical solution to the posterior was achieved by recognizing that Eq. (7) is algebraically equivalent to having a Gaussian "working likelihood" of the pseudo-outcomes $\hat{Y}_i^{\mathrm{AIPW}} \mid \theta \sim \mathcal{N}(\theta, \omega^{-1})$ and employing the normal-normal conjugacy. Under this interpretation, the generalized posterior is calibrated by setting $\omega^{-1} = \mathrm{Var}(\hat{Y}^{\mathrm{AIPW}})$. Since $\hat{\theta}^{\mathrm{AIPW}}$ is the efficient estimator of θ_{ATE}, it achieves the lowest possible asymptotic variance among all regular, asymptotically linear (RAL) estimators. Consequently,

$$s_p^2 \to (n\omega)^{-1} = n^{-1}\mathrm{Var}(\hat{Y}^{\mathrm{AIPW}}) = \mathrm{Var}(\hat{\theta}_{\mathrm{ATE}}), \tag{8}$$

which implies that, under a diffuse prior, the corresponding frequentist-calibrated generalized posterior achieves the smallest achievable frequentist-calibrated variance and has, in turn, the narrowest faithful credible intervals.

6 Experiments

In our experiments, we investigate how generalized posteriors perform across different causal estimation strategies, focusing on (i) faithful and efficient uncertainty quantification, (ii) robustness to model misspecification, and (iii) sensitivity to calibration and prior information. We present results for the back-door ATE setting, while CATE estimation and additional experiments are reported in Appendix D.

Strategy	Orthogonal	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$
RA	✗	0.580 (0.432-0.718)	0.520 (0.374-0.663)	0.480 (0.337-0.626)	0.680 (0.533-0.805)	0.580 (0.432-0.718)	0.080 (0.022-0.192)	0.300 (0.179-0.446)	0.460 (0.318-0.607)	0.420 (0.282-0.568)
IPW	✗	1.000 (0.929-1.000)	1.000 (0.929-1.000)	1.000 (0.929-1.000)	1.000 (0.929-1.000)	1.000 (0.929-1.000)	0.420 (0.282-0.568)	0.980 (0.894-0.999)	1.000 (0.929-1.000)	1.000 (0.929-1.000)
AIPW	✓	**0.940** (0.835-0.987)	**0.980** (0.894-0.999)	**0.980** (0.894-0.999)	**0.920** (0.808-0.978)	**0.940** (0.835-0.987)	**0.580** (0.432-0.718)	**0.940** (0.835-0.987)	**0.980** (0.894-0.999)	**0.940** (0.835-0.987)

Table 1: **Coverage for back-door ATE.** The 95% $\mathrm{CrI}^{(r)}_{0.95}$ credible interval's coverage $\hat{C}$ (and its 95% confidence interval) for each strategy $S \in \{\mathrm{AIPW}, \mathrm{IPW}, \mathrm{RA}\}$ across 9 synthetic datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_9\}$. Results are computed from $R = 50$ repetitions at sample size $n = 1000$. Closest to exact frequentist calibration at 95% is in **bold**.

Settings: We consider nine synthetic back-door settings $\{\mathcal{D}_1, \ldots, \mathcal{D}_9\}$, each with a known ground-truth ATE (see Appendix E.2.1). We first compare generalized posteriors constructed using three standard estimation strategies $S \in \{\mathrm{RA}, \mathrm{IPW}, \mathrm{AIPW}\}$, with their corresponding nuisance components estimated using flexible ML models. For each strategy and dataset, we repeat the posterior estimation pipeline over $R = 50$ random seeds at sample size $n = 1000$. The subsequent robustness and sensitivity experiments are described alongside their respective results, with implementation details provided in Appendix E.

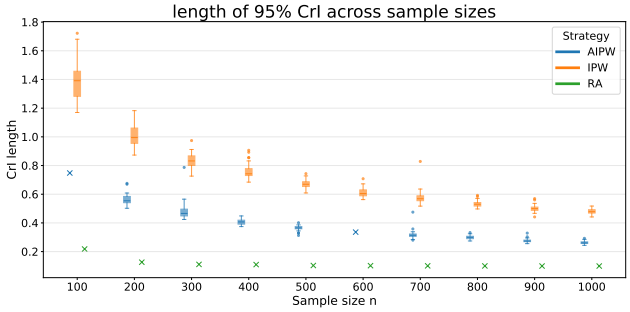

Figure 2: **CrI length for back-door ATE.** Box plots of the lengths of the 95% $\mathrm{CrI}^{(r)}_{0.95}$ credible interval across $R = 50$ repetitions for each strategy $S \in \{\mathrm{AIPW}, \mathrm{IPW}, \mathrm{RA}\}$, the dataset $\mathcal{D}_1$, and a varying sample size $n \in \{100, \ldots, 1000\}$. Unfaithful credible intervals are reported with a cross at their median length.

Metrics: We measure: (i) *Coverage:* For each repetition $r = 1, \ldots, R$, we construct a nominal 95% credible interval $\mathrm{CrI}^{(r)}_{0.95}$ and estimate its frequentist coverage as $\hat{C} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\{\theta^\star \in \mathrm{CrI}^{(r)}_{0.95}\}$. (ii) *Interval length:* We measure the average credible interval length across repetitions. To avoid favoring artificially narrow intervals, we mark as unfaithful those strategy-dataset pairs for which the upper bound of the 95% confidence interval for $\hat{C}$ is below 95%, and exclude them when comparing interval lengths. (iii) *Mean squared error (MSE):* For comparisons of point-estimation accuracy, we report the empirical MSE of the posterior mean relative to the true causal estimand $\theta^\star$.

Coverage and efficiency across estimation strategies: Table 1 presents the empirical coverage in each setting. Consistent with our theoretical results, the Neyman-orthogonal AIPW-based generalized posterior generally achieves the most faithful uncertainty quantification. Figure 2 further illustrates the efficiency of the AIPW-based generalized posterior as sample size increases. Detailed results on credible interval lengths are provided in Appendix D.2.

	Well-specified outcome model		Misspecified outcome model	
Method	Coverage $\hat{C}$	MSE	Coverage $\hat{C}$	MSE
Standard Bayes	0.932 (0.906-0.952)	0.0050	0.006 (0.001-0.017)	0.1572
Generalized Bayes (AIPW)	0.932 (0.906-0.952)	0.0053	0.968 (0.949-0.982)	0.0227

Table 2: **Standard vs. generalized Bayes under outcome misspecification.** Empirical coverage of nominal 95% credible intervals (with 95% confidence intervals) and posterior-mean MSE under well-specified and misspecified outcome models. Results are based on $R = 500$ repetitions with $n = 1000$, using a diffuse $\mathcal{N}(0, 100^2)$ prior.

Robustness to model misspecification: Table 2 compares likelihood-based Bayesian inference with our AIPW-based generalized posterior under correctly specified and partially misspecified outcome models. While both approaches perform similarly under correct specification, the generalized posterior retains near-nominal coverage under partial misspecification, illustrating how the double

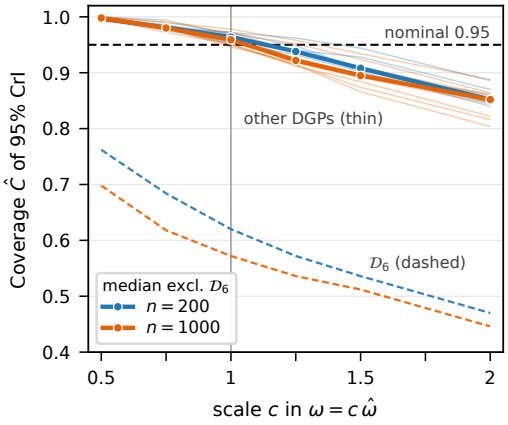

Figure 3: **Sensitivity to the calibration parameter ω.** Coverage $\hat{C}$ of the 95% credible interval of the AIPW generalized posterior when the plug-in calibration $\hat{\omega} = \widehat{\mathrm{Var}}(\hat{Y}^{\mathrm{AIPW}})^{-1}$ is rescaled to $\omega = c\,\hat{\omega}$, across $\{\mathcal{D}_1, \ldots, \mathcal{D}_9\}$ at $n \in \{200, 1000\}$ ($R = 500$ repetitions). Bold: median across datasets excluding $\mathcal{D}_6$; thin: individual datasets; dashed: $\mathcal{D}_6$ (limited overlap), which under-covers for every c.

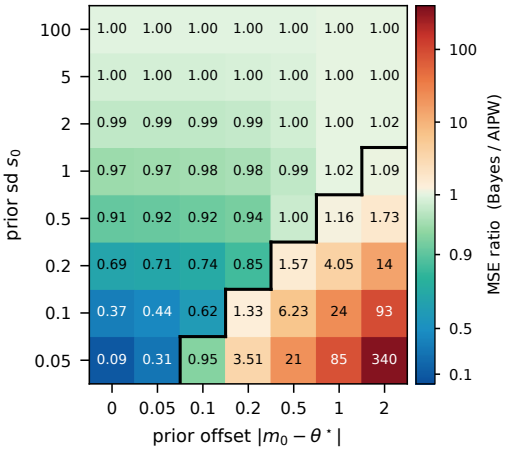

Figure 4: **Prior strength: when does the prior help?** Ratio of the MSE of the generalized posterior mean to the MSE of the frequentist $\hat{\theta}^{\mathrm{AIPW}}$ on the same data, for priors $\mathcal{N}(m_0, s_0^2)$ with offset $|m_0 - \theta^\star|$ and scale s_0 ($\mathcal{D}_1$, $n = 1000$, $R = 500$ per cell). Green: the prior improves MSE (ratio < 1); red: it hurts. The black line marks the asymptotic boundary $|m_0 - \theta^\star| = \sqrt{2}\,s_0$.

robustness of the AIPW loss translates into robust uncertainty quantification when the propensity model remains correctly specified.

Sensitivity to posterior calibration: Figure 3 examines the sensitivity of frequentist coverage to the choice of calibration parameter ω. The results highlight the importance of appropriate calibration for faithful uncertainty quantification, while also showing that calibration alone cannot remedy persistent undercoverage in challenging settings such as limited overlap (i.e. setting $\mathcal{D}_6$). Analytic and bootstrap calibration yield comparable results, with further details in Appendix D.3.

The role of prior information: To gain intuition about when prior information can improve causal effect estimation, consider a simple setting where the generalized posterior is available analytically. Suppose $\hat{\theta}$ is an unbiased estimator of $\theta^\star$ with variance v_n, and we place a Gaussian prior $\pi(\theta) = \mathcal{N}(m_0, s_0^2)$ directly on the causal estimand. With ω calibrated to v_n, we can show that the generalized posterior mean achieves lower frequentist MSE than $\hat{\theta}$ if and only if

$$(m_0 - \theta^\star)^2 < 2s_0^2 + v_n. \tag{9}$$

This illustrates how incorporating prior information can improve estimation accuracy when the prior is sufficiently well-centered, while a strongly informative but poorly centered prior may instead be harmful. Figure 4 illustrates this trade-off empirically across different prior specifications.

Conclusion: We have presented a generalized Bayesian framework for causal inference that places priors directly on causal estimands and updates them using identification-driven losses, without requiring a likelihood model for the full data-generating process. For Neyman-orthogonal losses, our theoretical results establish robustness to nuisance estimation error and asymptotically valid uncertainty quantification under suitable regularity and calibration conditions. Our experiments illustrate these properties, while also highlighting the importance of calibration and the challenges posed by limited overlap. Further experiments, including CATE estimation and sensitivity analyses, are reported in Appendix D, with additional practical considerations in Appendix B. The framework remains subject to causal identification assumptions and finite-sample limitations, discussed further in Appendix F.

Acknowledgments and Disclosure of Funding

This paper is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space. Additionally, this work has been supported by the German Federal Ministry of Education and Research (Grant: 01IS24082). We also thank Valentyn Melnychuk for valuable comments and suggestions on earlier versions of this manuscript.

References

[1] A. M. Alaa and M. van der Schaar. Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes. *arXiv preprint*, arXiv:1704.02801, 2017.

[2] J. Antonelli, G. Papadogeorgou, and F. Dominici. Causal Inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties. *Biometrics*, 78(1): 100–114, 2022.

[3] K. Babasaki, S. Sugasawa, K. Takanashi, and K. McAlinn. Ensemble Doubly Robust Bayesian Inference via Regression Synthesis. *arXiv preprint*, arXiv:2409.06288, 2024.

[4] C. R. S. Banerji, T. Chakraborti, C. Harbron, and B. D. MacArthur. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature Medicine*, 29(12):2996–2998, 2023.

[5] P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(5): 1103–1130, 2016.

[6] C. Breunig, R. Liu, and Z. Yu. Double robust Bayesian inference on average treatment effects. *Econometrica*, 93(2):539–568, 2025.

[7] V. Chernozhukov and H. Hong. An MCMC Approach to Classical Estimation. *Journal of Econometrics*, 115(2):293–346, 2003.

[8] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):1–68, 2018.

[9] S. Feuerriegel, D. Frauen, V. Melnychuk, J. Schweisthal, K. Hess, A. Curth, S. Bauer, N. Kilbertus, I. S. Kohane, and M. van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

[10] D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3), 2023.

[11] D. T. Frazier and D. J. Nott. Cutting Feedback and Modularized Analyses in Generalized Bayesian Inference. *Bayesian Analysis*, 20(4):1647–1675, 2025.

[12] D. Freedman. On the Bernstein-Von Mises Theorem with Infinite-Dimensional Parameters. *The Annals of Statistics*, 27(4):1119–1140, 1999.

[13] R. Gao, M. Deistler, and J. H. Macke. Generalized Bayesian Inference for Scientific Simulators via Amortized Cost Estimation. In *NeurIPS*, 2023.

[14] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4): 1360–1383, 2008.

[15] T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[16] G. Goh and J. Yu. Bayesian causal inference with some invalid instrumental variables. *Oxford Bulletin of Economics and Statistics*, 84(6):1432–1451, 2022.

[17] P. R. Hahn, C. M. Carvalho, D. Puelz, and J. He. Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis*, 13(1):163–182, 2018.

[18] J. J. Heckman, J. Smith, and N. Clements. Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.

[19] P. E. Jacob, L. M. Murray, C. C. Holmes, and C. P. Robert. Better together? Statistical learning in models made of modules. *arXiv preprint*, arXiv:1708.08719, 2017.

[20] E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

[21] E. H. Kennedy. Semiparametric Doubly Robust Targeted Double Machine Learning: A Review. In *Handbook of Statistical Methods for Precision Medicine*. Chapman and Hall/CRC, 2024.

[22] T. Kneib, A. Silbersdorff, and B. Säfken. Rage Against the Mean – A Review of Distributional Regression Approaches. *Econometrics and Statistics*, 26:99–123, 2023.

[23] J. Knoblauch. Frequentist Consistency of Generalized Variational Inference. *arXiv preprint*, arXiv:1912.04946, 2019.

[24] J. Knoblauch, J. Jewson, and T. Damoulas. Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β-Divergences. In *NeurIPS*, 2018.

[25] J. Knoblauch, J. Jewson, and T. Damoulas. An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference. *Journal of Machine Learning Research*, 23 (132):1–109, 2022.

[26] A. R. Linero and J. L. Antonelli. The How and Why of Bayesian Nonparametric Causal Inference. *WIREs Computational Statistics*, 15(1), 2023.

[27] T. Matsubara, J. Knoblauch, F.-X. Briol, and C. J. Oates. Robust Generalised Bayesian Inference for Intractable Likelihoods. *Journal of the Royal Statistical Society Series B*, 84(3):997–1022, 2022.

[28] J. W. Miller. Asymptotic Normality, Concentration, and Coverage of Generalized Posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

[29] I. D. Muñoz, A. E. Hubbard, and M. J. van der Laan. Targeted Bayesian Learning. In M. J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics. Springer, New York, 2011. ISBN 978-1-4419-9782-1.

[30] X. Nie and S. Wager. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika*, 108(2):299–319, 2021.

[31] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley, 1 edition, 2006. ISBN 978-0-470-02999-2 978-0-470-03331-9.

[32] S. Orihara, T. Momozaki, and S. Sugasawa. Bayesian Doubly Robust Causal Inference via Posterior Coupling. *arXiv preprint*, arXiv:2506.04868, 2025.

[33] K. Ray and A. van der Vaart. Semiparametric Bayesian causal inference. *The Annals of Statistics*, 48(5), 2020.

[34] J. M. Robins, M. A. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–60, 2000.

[35] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[36] O. Saarela, L. R. Belzile, and D. A. Stephens. A Bayesian view of doubly robust causal inference. *Biometrika*, 103(3):667–681, 2016.

[37] M. Sabbagh and D. A. Stephens. Bootstrap validity in Bayesian semi-parametric models. *arXiv preprint*, arXiv:2608.06670, 2026.

[38] M. Sabbagh and D. A. Stephens. Semi-parametric Bayesian inference under Neyman orthogonality. *arXiv preprint*, arXiv:2602.20371, 2026.

[39] H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

[40] G. Sert, A. Chakrabortty, and A. Bhattacharya. Bayesian Semiparametric Causal Inference: Targeted Doubly Robust Estimation of Treatment Effects. *arXiv preprint*, arXiv:2511.15904, 2025.

[41] D. A. Stephens, W. S. Nobre, E. E. M. Moodie, and A. M. Schmidt. Causal Inference Under Mis-Specification: Adjustment Based on the Propensity Score. *Bayesian Analysis*, 18(2): 639–694, 2023.

[42] N. Syring and R. Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2): 479–486, 2019.

[43] M. J. Van Der Laan. Statistical Inference for Variable Importance. *The International Journal of Biostatistics*, 2(1), 2006.

[44] M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer New York, New York, NY, 2011. ISBN 978-1-4419-9781-4.

[45] M. J. van der Laan and D. Rubin. Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1), 2006.

[46] Z. Wang, Y. Zhou, T. Ren, and J. Zhu. Quasi-Bayesian Dual Instrumental Variable Regression. In *NeurIPS*, 2021.

[47] A. Yiu, E. Fong, C. Holmes, and J. Rousseau. Semiparametric posterior corrections. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1025–1054, 2025.

[48] F. G. Zampieri, J. D. Casey, M. Shankar-Hari, F. E. Harrell, and M. O. Harhay. Using Bayesian Methods to Augment the Interpretation of Critical Care Trials. An Overview of Theory and Example Reanalysis of the Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial. *American Journal of Respiratory and Critical Care Medicine*, 203(5):543–552, 2021.

[49] W. Zhang, L. Liu, and J. Li. Treatment Effect Estimation with Disentangled Latent Factors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10923–10930, 2021.

A Proofs

We provide the formal statement and proof of Theorem 1. We consider (a) finite-dimensional θ, and (b) finite-dimensional projections of an infinite-dimensional θ. In the finite-dimensional case, we first derive the feasible Bernstein–von Mises (BvM) approximation from regularity conditions on the cross-fitted loss, then compare it with the oracle posterior.

A.1 Finite dimensional θ

Theorem 2 (Posterior stability under orthogonal losses (finite-dimensional θ))**.** *Let $\Theta \subset \mathbb{R}^d$ be open, and let $\eta \in \mathcal{H}$ be a nuisance parameter with truth η_0. For a loss $\ell(O;\theta,\eta)$, define the empirical and population risks*

$$\mathcal{L}_n(\theta;\eta_0) := \frac{1}{n}\sum_{i=1}^{n} \ell(O_i;\theta,\eta_0), \qquad L(\theta;\eta) := \mathbb{E}_{\mathbb{P}_{\rm obs}}\{\ell(O;\theta,\eta)\}. \tag{10}$$

Let $\theta^\star$ be the unique minimizer of $L(\cdot;\eta_0)$, and suppose that $V_0 := \nabla_{\theta\theta} L(\theta^\star;\eta_0)$ is positive definite. For a fixed K, partition the observations into validation folds $I_1,\dots,I_K$, and let $\hat\eta^{(-k)}$ be fitted without the observations in I_k. Define the cross-fitted feasible risk

$$\mathcal{L}_n(\theta;\hat\eta) := \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k} \ell(O_i;\theta,\hat\eta^{(-k)}). \tag{11}$$

Fix $\omega > 0$ and a proper prior density π on Θ. Define

$$q_{n,\rm or}(\theta \mid \mathcal{D}_n) \;\propto\; \pi(\theta)\exp\{-\omega n \mathcal{L}_n(\theta;\eta_0)\}, \qquad q_{n,\rm fe}(\theta \mid \mathcal{D}_n) \;\propto\; \pi(\theta)\exp\{-\omega n \mathcal{L}_n(\theta;\hat\eta)\}, \tag{12}$$

and their empirical minimizers

$$\hat\theta_{\rm or} := \arg\min_{\theta\in\Theta} \mathcal{L}_n(\theta;\eta_0), \qquad \hat\theta_{\rm fe} := \arg\min_{\theta\in\Theta} \mathcal{L}_n(\theta;\hat\eta). \tag{13}$$

Let $r_n \downarrow 0$ be a deterministic nuisance-rate sequence, and let $B_0 := B(\theta^\star,\delta_0)$ for a fixed $\delta_0 > 0$ with $\overline{B}_0 \subset \Theta$. Assume:

(A1) ***(Oracle BvM).*** *The oracle generalized posterior satisfies*

$$\mathrm{TV}\Big(q_{n,\rm or}(\cdot \mid \mathcal{D}_n),\; \mathcal{N}\big(\hat\theta_{\rm or}, (n\omega V_0)^{-1}\big)\Big) = o_{\mathbb{P}_{\rm obs}}(1). \tag{14}$$

This is the oracle BvM input; no BvM approximation for the feasible posterior is assumed.

(A2) ***(Neyman orthogonality).*** *The loss is Neyman orthogonal at $(\theta^\star,\eta_0)$ in the sense of Eq.* (4)*.*

(A3) ***(Orthogonal-score regularity).*** *Write $\psi(O;\theta,\eta) := \nabla_\theta \ell(O;\theta,\eta)$ and $\Delta\psi_k(O) := \psi(O;\theta^\star,\hat\eta^{(-k)}) - \psi(O;\theta^\star,\eta_0)$. Assume $\|\hat\theta_{\rm or} - \theta^\star\| = O_{\mathbb{P}_{\rm obs}}(n^{-1/2})$ and, uniformly over $k \in \{1,\dots,K\}$,*

$$\big\|\mathbb{E}_{\mathbb{P}_{\rm obs}}[\Delta\psi_k(O)]\big\| = O_{\mathbb{P}_{\rm obs}}(r_n^2), \qquad \mathbb{E}_{\mathbb{P}_{\rm obs}}\big[\|\Delta\psi_k(O)\|^2\big] = O_{\mathbb{P}_{\rm obs}}(r_n^2), \tag{15}$$

where the expectations are conditional on the corresponding fold's training sample. Further assume the local Hessian perturbation bound

$$\sup_{\theta\in B_0} \big\|\nabla_{\theta\theta}\mathcal{L}_n(\theta;\hat\eta) - \nabla_{\theta\theta}\mathcal{L}_n(\theta;\eta_0)\big\| = O_{\mathbb{P}_{\rm obs}}(r_n). \tag{16}$$

These are standard types of second-order bias, stochastic equicontinuity, and local smoothness controls used in orthogonal M-estimation; they make the center stability argument below explicit.

(A4) ***(Regularity of the cross-fitted feasible risk).*** *The following conditions hold for the same B_0:*

(C1) ***Prior regularity.*** *π is continuous and strictly positive at $\theta^\star$.*

(C2) ***Local smoothness.*** *For η in a neighborhood $\mathcal{N}$ of η_0, the map $\theta \mapsto \ell(O;\theta,\eta)$ is three times continuously differentiable on a neighborhood of $\overline{B}_0$. There are integrable envelopes $M_j(O)$ such that*

$$\sup_{\theta\in\overline{B}_0,\ \eta\in\mathcal{N}} \|\nabla_\theta^j \ell(O;\theta,\eta)\| \le M_j(O), \qquad \mathbb{E}_{\mathbb{P}_{\mathrm{obs}}}[M_j(O)] < \infty, \quad j = 0,1,2,3. \tag{17}$$

(C3) ***Cross-fitting and nuisance consistency.*** *K is fixed, $|I_k|/n \to \alpha_k \in (0,1)$, and $\hat{\eta}^{(-k)}$ uses only $\{O_i : i \notin I_k\}$. With probability approaching one, $\hat{\eta}^{(-k)} \in \mathcal{N}$ for each k, and*

$$\max_{1\le k\le K} \|\hat{\eta}^{(-k)} - \eta_0\| = O_{\mathbb{P}_{\mathrm{obs}}}(r_n), \qquad r_n \to 0. \tag{18}$$

(C4) ***Identification and local curvature.*** *$\theta^\star$ is the unique minimizer of $L(\cdot;\eta_0)$. For constants $\lambda_0, c_{\mathrm{sep}} > 0$,*

$$\begin{aligned} \nabla_{\theta\theta} L(\theta;\eta_0) &\succeq 2\lambda_0 I_d \qquad \text{for all } \theta \in \overline{B}_0, \\ \inf_{\theta\notin B_0} \big\{L(\theta;\eta_0) - L(\theta^\star;\eta_0)\big\} &\ge c_{\mathrm{sep}}. \end{aligned} \tag{19}$$

(C5) ***Continuity in the nuisance.*** *The maps $\eta \mapsto L(\theta;\eta)$ and $\eta \mapsto \nabla_{\theta\theta} L(\theta;\eta)$ are continuous at η_0 uniformly over $\theta \in \overline{B}_0$.*

(C6) ***Global empirical separation.*** *With probability approaching one,*

$$\inf_{\theta\notin B_0} \big\{\mathcal{L}_n(\theta;\hat{\eta}) - \mathcal{L}_n(\theta^\star;\hat{\eta})\big\} \ge \frac{c_{\mathrm{sep}}}{2}. \tag{20}$$

For example, this follows from (C4) and a suitable global uniform law of large numbers for the cross-fitted feasible risk. It is stated separately because a uniform law only on B_0 is insufficient to control posterior mass outside B_0.

Then the feasible generalized posterior admits the BvM approximation

$$\mathrm{TV}\Big(q_{n,\mathrm{fe}}(\cdot \mid \mathcal{D}_n),\ \mathcal{N}\big(\hat{\theta}_{\mathrm{fe}}, (n\omega V_0)^{-1}\big)\Big) = o_{\mathbb{P}_{\mathrm{obs}}}(1), \tag{21}$$

and the posterior-stability bound is

$$\mathrm{TV}(q_{n,\mathrm{fe}}(\cdot \mid \mathcal{D}_n),\ q_{n,\mathrm{or}}(\cdot \mid \mathcal{D}_n)) = O_{\mathbb{P}_{\mathrm{obs}}}\big(\sqrt{n}\, r_n^2\big) + o_{\mathbb{P}_{\mathrm{obs}}}(1). \tag{22}$$

In particular, if $r_n = o(n^{-1/4})$, then $\mathrm{TV}(q_{n,\mathrm{fe}}, q_{n,\mathrm{or}}) \to 0$ in probability and the feasible posterior inherits the oracle BvM limit.

Discussion of assumptions. Our assumptions draw on two established strands of theory: Neyman-orthogonal statistical learning [10], which motivates (A2)–(A3), and generalized-posterior Bernstein–von Mises (BvM) theory [28], which motivates (A1) and (A4). Specifically, (A1) provides the oracle BvM approximation as a benchmark, while (A2)–(A3) ensure that Neyman-orthogonality translates into second-order stability under nuisance estimation error, accounting also for the stochastic fluctuations introduced by estimated nuisances. Assumption (A4) provides sufficient regularity conditions to derive the feasible BvM approximation rather than assuming it directly. Conditions (C1)–(C5) ensure prior regularity, smoothness, nuisance consistency, and local curvature of the cross-fitted empirical risk, allowing for a local quadratic approximation of the feasible posterior. Condition (C6) additionally ensures global empirical separation, which is needed to control posterior mass outside the neighborhood where the local approximation holds. Together, these conditions allow us to establish posterior stability through a comparison of the oracle and feasible Gaussian approximations.

Proof of Theorem 2. We omit conditioning on $\mathcal{D}_n$ when it is unambiguous. The proof first establishes Eq. (21) from (C1)–(C6), then uses orthogonal-score stability to compare the posterior centers.

Step 1 (Local uniform laws for the cross-fitted risk). For $j \in \{0, 2\}$, write $\varphi_j(O; \theta, \eta) := \nabla_\theta^j \ell(O; \theta, \eta)$, with $\varphi_0 = \ell$. Fix a fold k, and let $\mathcal{T}_k$ be the sigma-field generated by its training data and any algorithmic randomness used for $\hat{\eta}^{(-k)}$. Conditional on $\mathcal{T}_k$, the observations in I_k are i.i.d. and independent of $\hat{\eta}^{(-k)}$. Using (C2) and a conditional uniform law of large numbers, for $j \in \{0, 2\}$,

$$\sup_{\theta \in B_0} \left\| \frac{1}{|I_k|} \sum_{i \in I_k} \varphi_j(O_i; \theta, \hat{\eta}^{(-k)}) - \mathbb{E}_{\mathbb{P}_{\text{obs}}}\big[\varphi_j(O; \theta, \hat{\eta}^{(-k)})\big] \right\| = o_{\mathbb{P}_{\text{obs}}}(1). \tag{23}$$

By (C3) and (C5), the corresponding expectations converge uniformly to their oracle counterparts. Summing over the fixed number of folds gives

$$\sup_{\theta \in B_0} \big|\mathcal{L}_n(\theta; \hat{\eta}) - L(\theta; \eta_0)\big| = o_{\mathbb{P}_{\text{obs}}}(1), \tag{24}$$

$$\sup_{\theta \in B_0} \big\|\nabla_{\theta\theta} \mathcal{L}_n(\theta; \hat{\eta}) - \nabla_{\theta\theta} L(\theta; \eta_0)\big\| = o_{\mathbb{P}_{\text{obs}}}(1). \tag{25}$$

Moreover, the third-derivative envelope in (C2) implies

$$\sup_{\theta \in B_0} \big\|\nabla_{\theta\theta\theta} \mathcal{L}_n(\theta; \hat{\eta})\big\| \le \frac{1}{n} \sum_{i=1}^{n} M_3(O_i) = O_{\mathbb{P}_{\text{obs}}}(1). \tag{26}$$

Step 2 (Consistency and local curvature of the feasible minimizer). The global separation condition (C6) places any global minimizer of $\mathcal{L}_n(\cdot; \hat{\eta})$ inside B_0 with probability tending to one. On B_0, Eq. (24), population identification in (C4), and the argmin theorem give $\hat{\theta}_{\text{fe}} \to \theta^\star$ in probability. By (C4) and Eq. (25), with probability tending to one,

$$\nabla_{\theta\theta} \mathcal{L}_n(\theta; \hat{\eta}) \succeq \lambda_0 I_d \qquad \text{for every } \theta \in B_0. \tag{27}$$

Also, the consistency of $\hat{\theta}_{\text{fe}}$ and Eq. (25) imply

$$\nabla_{\theta\theta} \mathcal{L}_n(\hat{\theta}_{\text{fe}}; \hat{\eta}) = V_0 + o_{\mathbb{P}_{\text{obs}}}(1). \tag{28}$$

Step 3 (Local quadratic expansion). Let $h := \sqrt{n}(\theta - \hat{\theta}_{\text{fe}})$ and $\theta(h) := \hat{\theta}_{\text{fe}} + h/\sqrt{n}$. Let Q_n be the distribution of h induced by $q_{n,\text{fe}}$ and set $\Phi := \mathcal{N}(0, (\omega V_0)^{-1})$. The unnormalized density of Q_n, up to a constant independent of h, is

$$\tilde{q}_n(h) := \pi(\theta(h)) \exp\Big\{-\omega n\big(\mathcal{L}_n(\theta(h); \hat{\eta}) - \mathcal{L}_n(\hat{\theta}_{\text{fe}}; \hat{\eta})\big)\Big\}. \tag{29}$$

Since the feasible minimizer is interior with high probability, $\nabla_\theta \mathcal{L}_n(\hat{\theta}_{\text{fe}}; \hat{\eta}) = 0$. Taylor expansion and Eq. (28) yield, for each fixed $M < \infty$,

$$\sup_{\|h\| \le M} \left| n\big(\mathcal{L}_n(\theta(h); \hat{\eta}) - \mathcal{L}_n(\hat{\theta}_{\text{fe}}; \hat{\eta})\big) - \frac{1}{2} h^\top V_0 h \right| = o_{\mathbb{P}_{\text{obs}}}(1). \tag{30}$$

By (C1) and $\hat{\theta}_{\text{fe}} \to \theta^\star$,

$$\sup_{\|h\| \le M} \left| \frac{\pi(\theta(h))}{\pi(\hat{\theta}_{\text{fe}})} - 1 \right| = o_{\mathbb{P}_{\text{obs}}}(1). \tag{31}$$

Step 4 (Posterior tail control). For clarity, write $\Delta\mathcal{L}_n(\theta) := \mathcal{L}_n(\theta; \hat{\eta}) - \mathcal{L}_n(\hat{\theta}_{\text{fe}}; \hat{\eta})$. Choose $\delta_1 > 0$ small enough that $\overline{B(\theta^\star, 3\delta_1)} \subset B_0$ and π is bounded on $B(\theta^\star, 3\delta_1)$. With high probability, $\hat{\theta}_{\text{fe}} \in B(\theta^\star, \delta_1)$. The curvature bound (27) then gives

$$\Delta\mathcal{L}_n(\theta) \ge \frac{\lambda_0}{2} \|\theta - \hat{\theta}_{\text{fe}}\|^2 \qquad \text{for all } \theta \in B_0. \tag{32}$$

The local expansion in Step 3 and positivity of π near $\theta^\star$ provide a lower bound for the normalizing integral:

$$\int_\Theta \pi(\theta) \exp\{-\omega n \Delta\mathcal{L}_n(\theta)\}\, d\theta \ge c\, n^{-d/2} \qquad \text{with probability tending to one,} \tag{33}$$

for a fixed $c > 0$.

For $M < \sqrt{n}\,\delta_1$, split the tail region $\{\|h\| > M\}$ into (i) points with $M/\sqrt{n} < \|\theta - \hat{\theta}_{\mathrm{fe}}\| \le \delta_1$, (ii) points in B_0 farther than δ_1 from the minimizer, and (iii) points outside B_0. In region (i), the boundedness of π and Eq. (32) yield a contribution to the normalizing integral bounded by

$$Cn^{-d/2} \int_{\|h\|>M} \exp\left\{-\frac{\omega\lambda_0}{2}\|h\|^2\right\} dh. \tag{34}$$

In region (ii), Eq. (32) provides an exponential bound $\exp(-\omega n\lambda_0\delta_1^2/2)$. In region (iii), (C6) and the minimizing property of $\hat{\theta}_{\mathrm{fe}}$ give $\Delta\mathcal{L}_n(\theta) \ge c_{\mathrm{sep}}/2$. Since π is a proper density, the total integrals over (ii) and (iii) are exponentially small. Dividing by Eq. (33), we conclude that, for constants $C, c' > 0$ and all sufficiently large n,

$$Q_n(\{\|h\| > M\}) \le C \int_{\|h\|>M} e^{-\omega\lambda_0\|h\|^2/2}\, dh + Cn^{d/2}e^{-c'n} \quad \text{with probability tending to one.} \tag{35}$$

Thus the local laws Q_n are tight in probability.

Step 5 (Feasible BvM). Write $g(h) := \exp\{-\omega h^\top V_0 h/2\}$ and let $\phi(h) := g(h)/\int_{\mathbb{R}^d} g(u)\,du$ be the density of Φ. Equations (30) and (31) imply

$$\sup_{\|h\|\le M} \left| \frac{\tilde{q}_n(h)}{\pi(\hat{\theta}_{\mathrm{fe}})g(h)} - 1 \right| = o_{\mathbb{P}_{\mathrm{obs}}}(1) \qquad \text{for every fixed } M < \infty. \tag{36}$$

Combining this uniform local approximation with Eq. (35) gives convergence of the normalized density of Q_n to ϕ in L^1, and hence

$$\mathrm{TV}\Big(q_{n,\mathrm{fe}}, \mathcal{N}\big(\hat{\theta}_{\mathrm{fe}}, (n\omega V_0)^{-1}\big)\Big) = o_{\mathbb{P}_{\mathrm{obs}}}(1), \tag{37}$$

which proves Eq. (21) without assuming it.

Step 6 (Orthogonality and the center shift). Set $R_n(\theta) := \nabla_\theta\mathcal{L}_n(\theta;\hat{\eta}) - \nabla_\theta\mathcal{L}_n(\theta;\eta_0)$. For the fold-specific score difference $\Delta\psi_k$ in (A3), conditional on the training sample the held-out observations are i.i.d. By the bias and second-moment bounds in Eq. (15) and fixed K,

$$\|R_n(\theta^\star)\| = O_{\mathbb{P}_{\mathrm{obs}}}\left(r_n^2 + \frac{r_n}{\sqrt{n}}\right). \tag{38}$$

The local Hessian bound (16) and $\|\hat{\theta}_{\mathrm{or}} - \theta^\star\| = O_{\mathbb{P}_{\mathrm{obs}}}(n^{-1/2})$ imply

$$\|R_n(\hat{\theta}_{\mathrm{or}}) - R_n(\theta^\star)\| = O_{\mathbb{P}_{\mathrm{obs}}}(r_n n^{-1/2}). \tag{39}$$

Both minimizers lie in B_0 with probability tending to one. Using the first-order conditions and the mean-value formula along the segment joining them, together with Eq. (27), gives

$$\|\hat{\theta}_{\mathrm{fe}} - \hat{\theta}_{\mathrm{or}}\| = O_{\mathbb{P}_{\mathrm{obs}}}\left(r_n^2 + \frac{r_n}{\sqrt{n}}\right). \tag{40}$$

The $r_n/\sqrt{n}$ term is the usual cross-fitted empirical fluctuation; it is negligible at the posterior scale since $r_n \to 0$.

Step 7 (Gaussian comparison and posterior stability). Define

$$G_{\mathrm{or}} := \mathcal{N}\big(\hat{\theta}_{\mathrm{or}}, (n\omega V_0)^{-1}\big), \qquad G_{\mathrm{fe}} := \mathcal{N}\big(\hat{\theta}_{\mathrm{fe}}, (n\omega V_0)^{-1}\big). \tag{41}$$

By the triangle inequality,

$$\mathrm{TV}(q_{n,\mathrm{fe}}, q_{n,\mathrm{or}}) \le \mathrm{TV}(q_{n,\mathrm{fe}}, G_{\mathrm{fe}}) + \mathrm{TV}(G_{\mathrm{fe}}, G_{\mathrm{or}}) + \mathrm{TV}(G_{\mathrm{or}}, q_{n,\mathrm{or}}). \tag{42}$$

The first and last terms are $o_{\mathbb{P}_{\mathrm{obs}}}(1)$ by Eqs. (37) and (14). For the middle term, Pinsker's inequality and the KL formula for Gaussian distributions with common covariance give

$$\begin{aligned} \mathrm{TV}(G_{\mathrm{fe}}, G_{\mathrm{or}}) &\le \frac{1}{2}\sqrt{n\omega\,(\hat{\theta}_{\mathrm{fe}} - \hat{\theta}_{\mathrm{or}})^\top V_0 (\hat{\theta}_{\mathrm{fe}} - \hat{\theta}_{\mathrm{or}})} \\ &\le \frac{1}{2}\sqrt{n\omega\,\lambda_{\max}(V_0)}\,\|\hat{\theta}_{\mathrm{fe}} - \hat{\theta}_{\mathrm{or}}\|. \end{aligned} \tag{43}$$

Therefore, by Eq. (40),

$$\mathrm{TV}(G_{\mathrm{fe}}, G_{\mathrm{or}}) = O_{\mathbb{P}_{\mathrm{obs}}}(\sqrt{n}\, r_n^2 + r_n). \tag{44}$$

Combining with Eq. (42) and using $r_n \to 0$ gives

$$\mathrm{TV}(q_{n,\mathrm{fe}}, q_{n,\mathrm{or}}) = O_{\mathbb{P}_{\mathrm{obs}}}(\sqrt{n}\, r_n^2) + o_{\mathbb{P}_{\mathrm{obs}}}(1), \tag{45}$$

which is Eq. (22). If $r_n = o(n^{-1/4})$, then $\sqrt{n}\, r_n^2 \to 0$, and the feasible posterior inherits the oracle BvM limit. □

A.2 Finite dimensional projections of an infinite dimensional θ

Theorem 3 (Posterior stability under orthogonal losses for infinite-dimensional θ (finite-dimensional projections))**.** *Let Θ be a separable Hilbert space with norm $\|\cdot\|_\Theta$. Let $\eta \in \mathcal{H}$ be a nuisance parameter with truth η_0, and define*

$$\mathcal{L}_n(\theta;\eta) := \frac{1}{n}\sum_{i=1}^{n} \ell(O_i;\theta,\eta), \qquad L(\theta;\eta) := \mathbb{E}_{\mathbb{P}_{\mathrm{obs}}}\{\ell(O;\theta,\eta)\}, \tag{46}$$

where $\mathbb{P}_{\mathrm{obs}}$ is the true observational distribution. Define the oracle and feasible generalized posteriors

$$q_{n,\mathrm{or}}(\theta \mid \mathcal{D}_n) \;\propto\; \pi(\theta)\exp\{-\omega n \mathcal{L}_n(\theta;\eta_0)\}, \qquad q_{n,\mathrm{fe}}(\theta \mid \mathcal{D}_n) \;\propto\; \pi(\theta)\exp\{-\omega n \mathcal{L}_n(\theta;\hat\eta)\}. \tag{47}$$

Let $\hat\theta_{\mathrm{or}} := \arg\min_{\theta\in\Theta} \mathcal{L}_n(\theta;\eta_0)$ and $\hat\theta_{\mathrm{fe}} := \arg\min_{\theta\in\Theta} \mathcal{L}_n(\theta;\hat\eta)$.

Fix $m \in \mathbb{N}$ and bounded linear functionals $T_1,\dots,T_m \in \Theta^*$, and write $T(\theta) := (T_1(\theta),\dots,T_m(\theta))^\top \in \mathbb{R}^m$. Let $q^T_{n,\mathrm{or}}$ and $q^T_{n,\mathrm{fe}}$ denote the induced (pushforward) posteriors on $\mathbb{R}^m$:

$$q^T_{n,\mathrm{or}} := q_{n,\mathrm{or}} \circ T^{-1}, \qquad q^T_{n,\mathrm{fe}} := q_{n,\mathrm{fe}} \circ T^{-1}. \tag{48}$$

Assume:

(B1) ***(Projection BvM, oracle, and feasible).*** *There exists a bounded, self-adjoint, strictly positive operator $V_0 := \nabla_{\theta\theta} L(\theta^\star;\eta_0)$ on Θ such that*

$$\mathrm{TV}\Big(q^T_{n,\mathrm{or}}(\cdot \mid \mathcal{D}_n),\ \mathcal{N}_m\big(T(\hat\theta_{\mathrm{or}}),\ (n\omega\, TV_0T^\top)^{-1}\big)\Big) = o_{\mathbb{P}_{\mathrm{obs}}}(1), \tag{49}$$

$$\mathrm{TV}\Big(q^T_{n,\mathrm{fe}}(\cdot \mid \mathcal{D}_n),\ \mathcal{N}_m\big(T(\hat\theta_{\mathrm{fe}}),\ (n\omega\, TV_0T^\top)^{-1}\big)\Big) = o_{\mathbb{P}_{\mathrm{obs}}}(1), \tag{50}$$

where $TV_0T^\top$ is the $m\times m$ matrix with entries $(TV_0T^\top)_{jk} := \langle T_j^\sharp,\ V_0T_k^\sharp\rangle_\Theta$, and $T_j^\sharp \in \Theta$ is the Riesz representer of T_j.

(B2) ***(Orthogonality).*** *The loss is Neyman orthogonal at $(\theta^\star,\eta_0)$ in the sense of Eq.* (4)*.*

(B3) ***(Cross-fitting and projected center stability).*** *$\hat\eta$ is obtained by sample splitting/cross-fitting with $\|\hat\eta - \eta_0\| =: r_n \to 0$. Assume additionally that the orthogonal M-estimation regularity conditions yield the projected center-shift bound $\|T(\hat\theta_{\mathrm{fe}}) - T(\hat\theta_{\mathrm{or}})\|_2 = O_{\mathbb{P}_{\mathrm{obs}}}(r_n^2)$.*

Then

$$\mathrm{TV}\big(q^T_{n,\mathrm{fe}}(\cdot \mid \mathcal{D}_n),\ q^T_{n,\mathrm{or}}(\cdot \mid \mathcal{D}_n)\big) \;=\; O_{\mathbb{P}_{\mathrm{obs}}}\big(\sqrt{n}\, r_n^2\big) \;+\; o_{\mathbb{P}_{\mathrm{obs}}}(1). \tag{51}$$

In particular, if $r_n = o(n^{-1/4})$, then $\mathrm{TV}(q^T_{n,\mathrm{fe}}, q^T_{n,\mathrm{or}}) \to 0$ in probability, so the feasible posterior inherits the oracle projection BvM limit from Eq. (49) *for $T(\theta)$.*

Proof. We omit the conditioning on $\mathcal{D}_n$ for readability.

Step 1 (Gaussian approximations in $\mathbb{R}^m$). By Eq. (49)–(50), define the Gaussian approximations

$$G^T_{\mathrm{or}} := \mathcal{N}_m\big(T(\hat\theta_{\mathrm{or}}),\ (n\omega\, TV_0T^\top)^{-1}\big), \qquad G^T_{\mathrm{fe}} := \mathcal{N}_m\big(T(\hat\theta_{\mathrm{fe}}),\ (n\omega\, TV_0T^\top)^{-1}\big), \tag{52}$$

so that $\mathrm{TV}(q^T_{n,\mathrm{or}}, G^T_{\mathrm{or}}) = o_{\mathbb{P}_{\mathrm{obs}}}(1)$ and $\mathrm{TV}(q^T_{n,\mathrm{fe}}, G^T_{\mathrm{fe}}) = o_{\mathbb{P}_{\mathrm{obs}}}(1)$.

Step 2 (Orthogonality implies a second-order center shift). Assumption (B3) supplies the projected center-shift bound

$$\|T(\hat{\theta}_{\mathrm{fe}}) - T(\hat{\theta}_{\mathrm{or}})\|_2 \;=\; O_{\mathbb{P}_{\mathrm{obs}}}(r_n^2). \tag{53}$$

This is the projected counterpart of the finite-dimensional M-estimation stability argument; it is assumed here rather than derived from the Hilbert-space conditions.

Step 3 (Triangle inequality).

$$\mathrm{TV}(q_{n,\mathrm{fe}}^T, q_{n,\mathrm{or}}^T) \le \mathrm{TV}(q_{n,\mathrm{fe}}^T, G_{\mathrm{fe}}^T) + \mathrm{TV}(G_{\mathrm{fe}}^T, G_{\mathrm{or}}^T) + \mathrm{TV}(G_{\mathrm{or}}^T, q_{n,\mathrm{or}}^T). \tag{54}$$

The first and third terms are $o_{\mathbb{P}_{\mathrm{obs}}}(1)$. It remains to bound $\mathrm{TV}(G_{\mathrm{fe}}^T, G_{\mathrm{or}}^T)$.

Step 4 (TV bound between Gaussians with common covariance). Both Gaussians have the same covariance $\Sigma_T = (n\omega\, TV_0T^\top)^{-1}$ and means $\mu_{\mathrm{fe}} = T(\hat{\theta}_{\mathrm{fe}})$, $\mu_{\mathrm{or}} = T(\hat{\theta}_{\mathrm{or}})$. For multivariate normals with equal covariance,

$$\mathrm{KL}(G_{\mathrm{fe}}^T \| G_{\mathrm{or}}^T) = \frac{1}{2}(\mu_{\mathrm{fe}} - \mu_{\mathrm{or}})^\top \Sigma_T^{-1}(\mu_{\mathrm{fe}} - \mu_{\mathrm{or}}) = \frac{n\omega}{2}(\mu_{\mathrm{fe}} - \mu_{\mathrm{or}})^\top (TV_0T^\top)(\mu_{\mathrm{fe}} - \mu_{\mathrm{or}}). \tag{55}$$

By Pinsker's inequality,

$$\mathrm{TV}(G_{\mathrm{fe}}^T, G_{\mathrm{or}}^T) \le \sqrt{\frac{1}{2}\,\mathrm{KL}(G_{\mathrm{fe}}^T \| G_{\mathrm{or}}^T)} \;\le\; \frac{1}{2}\sqrt{n\omega\,\lambda_{\max}(TV_0T^\top)}\;\|\mu_{\mathrm{fe}} - \mu_{\mathrm{or}}\|_2. \tag{56}$$

Using Eq. (53) and that $\lambda_{\max}(TV_0T^\top) = O(1)$ for fixed T, we get

$$\mathrm{TV}(G_{\mathrm{fe}}^T, G_{\mathrm{or}}^T) = O_{\mathbb{P}_{\mathrm{obs}}}(\sqrt{n}\, r_n^2). \tag{57}$$

Step 5 (Combine). Plugging into Eq. (54) yields $\mathrm{TV}(q_{n,\mathrm{fe}}^T, q_{n,\mathrm{or}}^T) = O_{\mathbb{P}_{\mathrm{obs}}}(\sqrt{n}\, r_n^2) + o_{\mathbb{P}_{\mathrm{obs}}}(1)$. If $r_n = o(n^{-1/4})$, then $\sqrt{n}\, r_n^2 \to 0$ and, hence, $\mathrm{TV}(q_{n,\mathrm{fe}}^T, q_{n,\mathrm{or}}^T) \to 0$ in probability. □

Remark 2 (CATE as a special case)**.** *If $\theta = \tau$ is the CATE function and T is the evaluation map $T(\tau) = (\tau(x_1), \ldots, \tau(x_m))$, the theorem gives stability (and inherited BvM) for any fixed finite set of evaluation points $x_{1:m}$, provided point evaluation is continuous in the chosen function space Θ.*

B Additional Details

B.1 Interpretation of generalized posteriors

Our framework constructs a generalized (Gibbs) posterior over the causal estimand rather than a conventional likelihood-based Bayesian posterior. The distinction is that, while the generalized posterior is a normalized probability distribution over the estimand, the exponentiated loss replacing the likelihood need not constitute a conditional probability distribution over the observed data. Consequently, the generalized posterior need not arise by conditioning a joint probabilistic model of the data and estimand with the specified prior as its marginal. This does not imply that generalized Bayesian updating is incoherent: for additive losses and a fixed learning rate, sequential updating yields the same generalized posterior as updating on the combined data. Moreover, specific losses may admit a working-likelihood interpretation, as illustrated by our Gaussian pseudo-outcome example in Section 5.4. Such a representation, however, does not amount to modeling the full observational data-generating process. Hence, the resulting uncertainty is interpreted as generalized-posterior uncertainty, with frequentist coverage requiring appropriate calibration and regularity conditions. We refer to [25] for a principled exposition on posterior generalizations.

B.2 Nuisance estimation

The generalized posterior is constructed using a loss that depends on plug-in nuisance estimates $\hat{\eta}^S$, without restricting how these estimates are obtained. In particular, a Bayesian nuisance model can be fitted on each training fold and its posterior mean used as the corresponding plug-in estimator. This includes Bayesian nonparametric models, such as Gaussian processes. Our theoretical guarantees

depend on the convergence rate of $\hat{\eta}^S$, rather than on the estimation procedure itself. Thus, Bayesian nuisance estimators are directly admissible provided their plug-in estimates satisfy the nuisance-rate and regularity conditions of Theorem 2.

Why not integrate over the nuisance posterior? An alternative to plug-in estimation is to explicitly propagate nuisance uncertainty by marginalizing over a posterior distribution $\Pi(\eta^S \mid D_n)$, yielding

$$\widetilde{q}_n^S(\theta \mid D_n) = \int q_n^S(\theta \mid D_n, \eta^S)\, \Pi(d\eta^S \mid D_n).$$

Our choice to employ plug-in nuisances instead has two motivations. First, our construction is modular: information from nuisance estimation enters the estimand-level update, but the resulting generalized posterior does not feed back into nuisance estimation. This one-way information flow is closely related to the motivation for cut-posterior approaches [19, 11], which seek to avoid potentially harmful feedback between model components.

Second, Neyman-orthogonality makes explicit propagation of nuisance uncertainty unnecessary for the first-order frequentist guarantees established in our framework. Under the conditions of Theorem 2, orthogonality removes the first-order sensitivity of the estimating equation to nuisance perturbations, such that the resulting posterior displacement is of order $O_P(r_n^2)$. When $r_n = o(n^{-1/4})$, this displacement is negligible relative to the $n^{-1/2}$ posterior scale. This parallels orthogonal statistical learning, where the asymptotic variance is determined by the oracle influence function without an additional first-order contribution from nuisance estimation [8]. Consequently, explicitly integrating over the nuisance posterior is not required for the asymptotic uncertainty guarantees considered here, while doing so would require specifying and computing a potentially high-dimensional nuisance posterior. Cross-fitting ensures that the held-out observations are conditionally i.i.d. within each fold given the corresponding nuisance estimates, providing the independence structure used in our theoretical analysis of posterior stability under Neyman-orthogonality.

B.3 Prior elicitation

A practical advantage of our framework is that prior beliefs can be specified directly on the causal estimand θ, rather than indirectly through a model of the observational distribution and its nuisance components. For scalar estimands such as the ATE, prior information may come from historical studies or meta-analyses [39], or from domain experts whose beliefs about plausible treatment effects can be translated into prior distributions using established elicitation procedures [31]. When substantive prior information is limited, weakly informative priors offer a practical alternative, with their scale chosen to reflect plausible effect sizes [14]. Importantly, we do not advocate noninformative priors as a general recommendation: prior specification should reflect the scientific context, available evidence, and scale of the estimand. In particular, when prior information is uncertain, sensitivity to reasonable alternative prior specifications should be assessed.

B.4 Choice of cross-fitting folds

The number of cross-fitting folds K governs a trade-off between nuisance estimation quality and computational cost. With K folds, each nuisance estimator is trained on a fraction $(K-1)/K$ of the data, so increasing K provides larger training samples but requires more nuisance model fits. Our theoretical guarantees assume a fixed K as $n \to \infty$, but do not prescribe an optimal choice. In practice, common choices are $K \in \{2, 5, 10\}$ [8], with $K = 5$ providing a reasonable default. Smaller values may be preferable for computationally expensive nuisance learners, while larger values can be beneficial when nuisance estimation is sensitive to training sample size. Provided the required nuisance convergence rates and regularity conditions hold, the asymptotic guarantees remain valid for any fixed K. In finite samples, sensitivity to the choice of K can be assessed empirically (see Appendix D.5).

C Extended Related Work

Generalized Bayesian inference: Related generalization ideas appear in econometrics under the terms of quasi-/Laplace posterior [7], and more recent optimization-centric formulations [25] give even broader formalizations of the same concept. Miller [28] provide general sufficient conditions for concentration, asymptotic normality (Bernstein-von Mises), and frequentist coverage of generalized posteriors, while Syring and Martin [42] develop a calibration procedure to achieve nominal frequentist coverage of the generalized posterior' credible sets.

Likelihood-based Bayesian causal inference: A large body of Bayesian causal inference proceeds by specifying a probabilistic model for (parts of) the observed data distribution and performing posterior inference under this model. Many approaches leverage flexible Bayesian nonparametric priors over high-dimensional nuisance components [e.g., 2, 33, 29]; see Linero and Antonelli [26] for an overview.

Comparison to Jacob et al. [19] and Frazier and Nott [11]: While Jacob et al. [19] and, in the generalized Bayes setting, Frazier and Nott [11] study modular inference for fully specified model modules, our cut structure is realized through an identification-driven Neyman-orthogonal loss, which is what delivers robustness to nuisance estimation error (Theorem 2).

Comparison to Yiu et al. [47]: The approach of Yiu et al. [47] starts from a conventional nonparametric Bayesian posterior over the full data-generating distribution and post-processes the marginal posterior of each target functional with a Bayesian-bootstrap one-step correction, yielding a corrected posterior that satisfies a Bernstein–von Mises (BvM) theorem with semiparametric-efficient width; the causal example developed there is the ATT. The difference to our work is along four axes:

1. *Where the prior lives.* Yiu et al. require a prior and posterior on P_{obs}; the marginal on θ remains a pushforward $\bar{Q}_{\#}\pi$, so there is no direct control over the induced prior on the estimand: precisely the difficulty we describe as point (4) of Section 4. Our framework never forms a posterior over P_{obs} and places the prior on θ directly, thus resolving this difficulty by construction.
2. *Where the debiasing enters.* They correct posterior draws post hoc using the efficient influence function. We build robustness into the update itself, through the identification-driven orthogonal loss. Consequently, any loss from the causal ML literature (e.g., DR-learner, R-learner) can be used as-is, without deriving and implementing an estimand-specific correction step.
3. *Scope of estimands.* Their construction targets low-dimensional functionals admitting an efficient influence function. Ours applies to any identified estimand expressible as an M-estimation problem, including infinite-dimensional targets such as the CATE function, where our guarantee is stated for finite-dimensional projections (Theorem 3) and where a one-step correction of a marginal posterior is not the natural object.
4. *What is required in practice.* Their method attaches to a posterior sampling algorithm for the full nonparametric model. Ours requires only a fitted nuisance learner and a loss, which is easier as the components are already present in a standard causal ML pipeline.

Concurrent work on Bayesian semiparametric inference: In concurrent work, Sabbagh and Stephens [38] study Bayesian semiparametric inference for target parameters defined through estimating equations involving estimated nuisance functions. Using a nonparametric Bayesian formulation based on the Dirichlet process and Bayesian bootstrap, they investigate when two-step plug-in inference remains valid despite not explicitly propagating posterior uncertainty about the nuisance functions. They show how Neyman-orthogonality justifies this approach and, interestingly, identify settings in which the posterior is asymptotically unaffected by nuisance estimation even without orthogonality, provided the nuisance estimator is consistent. In a follow-up paper, Sabbagh and Stephens [37] extend the analysis to sample splitting and cross-fitting, relaxing Donsker-type restrictions and establishing asymptotic properties of Bayesian bootstrap posteriors with flexible nuisance estimators. Together, these works provide a complementary perspective on Bayesian inference with estimated nuisances, highlighting the interplay between orthogonal estimating equations, posterior uncertainty, and the Bayesian bootstrap.

D Additional Experiments

D.1 CATE estimation

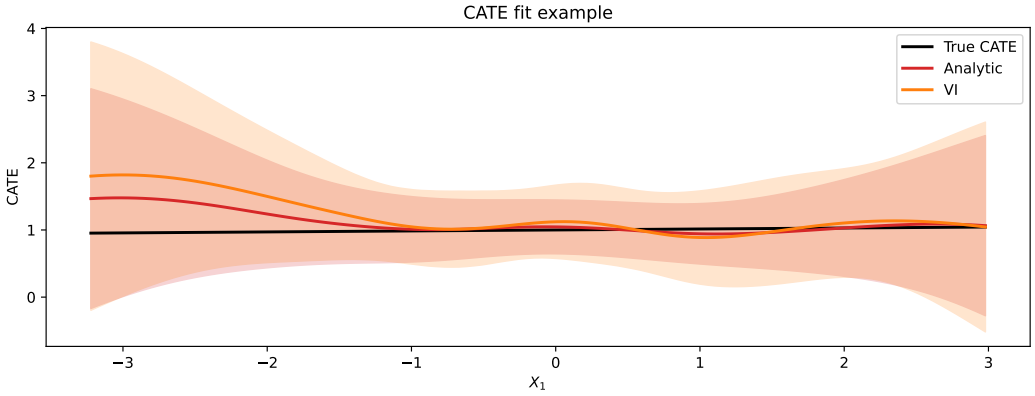

Figure 5: **(back-door CATE)** An example of a Gaussian Process fit (mean and 95% CrI) of the generalized posterior of CATE, with the DR loss, in $\mathcal{D}_4$, at $n = 1000$. The solution fitted with variational inference (VI) uses an Inducing Point GP. For implementation details, see Appendix E.

In Figure 5 we show an example of a fit of the generalized posterior for CATE. We report the coverage and length of credible interval for the back-door CATE setting in Tables 3 and 4. Note that the uncertainty around the *function* CATE is not a single interval, but rather a band around the mean fit. We report coverage and length metrics by averaging the point-wise uncertainty at $K = 100$ randomly sampled covariates X sampled from the underlying data-generating process.

Note that the CATE results are heavily influenced by the choice of kernel, which we have not optimized (for our implementation details, see Appendix E). Our aim here is to demonstrate the applicability of our approach to different settings and to show it can achieve frequentist-calibrated uncertainty coverage.

Strategy	Orthogonal	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$
RA	✗	**1.000** (0.964-1.000)	1.000 (0.964-1.000)	**1.000** (0.964-1.000)	**1.000** (0.964-1.000)	**1.000** (0.964-1.000)	0.500 (0.398-0.602)	0.760 (0.664-0.840)	0.420 (0.322-0.523)	0.000 (0.000-0.036)
IPW	✗	**1.000** (0.964-1.000)	0.900 (0.824-0.951)	**1.000** (0.964-1.000)	0.820 (0.731-0.890)	0.890 (0.812-0.944)	**1.000** (0.964-1.000)	**0.790** (0.697-0.865)	**1.000** (0.964-1.000)	**0.860** (0.776-0.921)
DR	✓	**1.000** (0.964-1.000)	**0.990** (0.946-1.000)	**1.000** (0.964-1.000)	**1.000** (0.964-1.000)	**1.000** (0.964-1.000)	0.800 (0.708-0.873)	0.770 (0.675-0.848)	**1.000** (0.964-1.000)	0.540 (0.437-0.640)

Table 3: **(back-door CATE)** The 95% $\mathrm{CrI}^{(r)}_{0.95}$ credible interval's coverage $\hat{C}$ (and its 95% confidence interval) for each strategy $S \in \{\mathrm{DR}, \mathrm{IPW}, \mathrm{RA}\}$ across 9 synthetic datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_9\}$. Results are computed from $R = 50$ repetitions at sample size $n = 1000$. Closest to exact frequentist calibration at 95% is in **bold** We report coverage metrics by averaging the point-wise uncertainty at $K = 100$ randomly sampled covariates X sampled from the underlying data-generating process. .

Strategy	Orthogonal	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$
RA	✗	**1.292 (0.000)**	**0.939 (0.000)**	**0.801 (0.000)**	1.606(0.000)	1.490(0.000)	1.505(0.000)	0.985(0.000)	3.563(0.000)	0.436(0.000)
IPW	✗	2.122(0.000)	1.965(0.000)	2.318(0.000)	3.718(0.000)	1.971(0.000)	**3.417 (0.000)**	2.361(0.000)	**5.484 (0.000)**	3.016(0.000)
DR	✓	1.378(0.000)	1.153(0.000)	1.524(0.000)	**1.554 (0.000)**	**1.321 (0.000)**	1.435(0.000)	2.308(0.000)	5.556(0.000)	0.932(0.000)

Table 4: **(back-door CATE)** The average length (and standard deviation) of the 95% $\mathrm{CrI}^{(r)}_{0.95}$ credible interval across $R = 50$ repetitions for each strategy $S \in \{\mathrm{DR}, \mathrm{IPW}, \mathrm{RA}\}$ across 9 synthetic datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_9\}$. Results are computed at sample size $n = 1000$. The lengths of unfaithful credible intervals are striked out in gray. The narrowest $\mathrm{CrI}^{(r)}_{0.95}$ is length is highlighted in **bold**. We report length metrics by averaging the point-wise uncertainty at $K = 100$ randomly sampled covariates X sampled from the underlying data-generating process.

D.2 Credible interval lengths for ATE estimation

Table 5 reports the credible interval lengths corresponding to the ATE coverage results in Table 1. Among methods achieving faithful coverage, the AIPW-based generalized posterior generally produces the shortest intervals, consistent with its theoretical efficiency properties.

D.3 Comparison of calibration methods

To complement the sensitivity analysis in the main text, we compare the uncalibrated choice $\omega = 1$ with analytic plug-in and bootstrap calibration across the nine synthetic settings, at $n \in \{200, 1000\}$.

Strategy	Orthogonal	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$
RA	✗	0.099(0.002)	0.112(0.004)	0.113(0.004)	0.139(0.007)	0.099(0.002)	0.112(0.009)	0.108(0.009)	0.108(0.002)	0.156(0.012)
IPW	✗	0.479(0.018)	0.490(0.019)	0.542(0.024)	0.687(0.067)	0.649(0.220)	0.693(0.256)	0.730(0.246)	0.644(0.222)	2.835(0.055)
AIPW	✓	**0.263 (0.010)**	**0.273 (0.009)**	**0.272 (0.008)**	**0.289 (0.009)**	**0.275 (0.021)**	0.446(0.136)	**0.587 (0.163)**	**0.336 (0.058)**	**0.466 (0.012)**

Table 5: **CrI length for back-door ATE.** The average length (and standard deviation) of the 95% $\mathrm{CrI}^{(r)}_{0.95}$ credible interval across $R = 50$ repetitions for each strategy $S \in \{\mathrm{AIPW}, \mathrm{IPW}, \mathrm{RA}\}$ across 9 synthetic datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_9\}$. Results are computed at sample size $n = 1000$. The lengths of unfaithful credible intervals are striked out in gray. The narrowest $\mathrm{CrI}^{(r)}_{0.95}$ is length is highlighted in **bold**.

Tables 6 and 7 show that both calibration methods achieve comparable, generally faithful coverage, with the analytic plug-in being computationally cheaper. Both substantially improve upon the uncalibrated posterior, although neither resolves persistent undercoverage in the limited-overlap setting $\mathcal{D}_6$.

n	Calibration	Mean cov. (excl. $\mathcal{D}_6$)	Range (excl. $\mathcal{D}_6$)	Cov. $\mathcal{D}_6$	Time [ms]
200	$\omega = 1$ (uncalibrated)	0.377	0.048–0.518	0.174	–
200	Analytic plug-in $\hat{\omega}$	0.963	0.954–0.972	0.620	0.08
200	Bootstrap $\hat{\omega}$	0.962	0.944–0.972	0.612	0.97
1000	$\omega = 1$ (uncalibrated)	0.494	0.308–0.630	0.128	–
1000	Analytic plug-in $\hat{\omega}$	0.960	0.946–0.978	0.572	0.08
1000	Bootstrap $\hat{\omega}$	0.955	0.930–0.980	0.564	0.97

Table 6: **Calibration regimes.** Coverage $\hat{C}$ of the 95% credible interval of the AIPW generalized posterior with an uncalibrated $\omega = 1$, the plug-in $\hat{\omega} = \widehat{\mathrm{Var}}(\hat{Y}^{\mathrm{AIPW}})^{-1}$, and a bootstrap $\hat{\omega} = (n\,\widehat{\mathrm{Var}}_{\mathrm{boot}}(\hat{\theta}^{\mathrm{AIPW}}))^{-1}$ computed from 100 nonparametric bootstrap resamples of the cross-fitted pseudo-outcomes. We report the mean and range over $\{\mathcal{D}_1, \ldots, \mathcal{D}_9\} \setminus \{\mathcal{D}_6\}$, the coverage on $\mathcal{D}_6$, and the median wall-clock time of the calibration step. Results are computed from $R = 500$ repetitions per dataset under a diffuse prior $\mathcal{N}(0, 100^2)$.

Calibration	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$
$\omega = 1$ (uncalibrated)	0.610 (0.566-0.653)	0.592 (0.547-0.635)	0.630 (0.586-0.672)	0.548 (0.503-0.592)	0.510 (0.465-0.555)	0.128 (0.100-0.160)	0.308 (0.268-0.351)	0.338 (0.297-0.381)	0.418 (0.374-0.463)
Analytic plug-in $\hat{\omega}$	0.946 (0.922-0.964)	0.966 (0.946-0.980)	0.972 (0.953-0.985)	**0.950** (0.927-0.967)	**0.950** (0.927-0.967)	**0.572** (0.527-0.616)	**0.952** (0.929-0.969)	**0.970** (0.951-0.983)	**0.978** (0.961-0.989)
Bootstrap $\hat{\omega}$	**0.952** (0.929-0.969)	**0.956** (0.934-0.972)	**0.968** (0.949-0.982)	0.938 (0.913-0.957)	0.944 (0.920-0.962)	0.564 (0.519-0.608)	0.930 (0.904-0.951)	0.970 (0.951-0.983)	0.980 (0.964-0.990)

Table 7: **Coverage by calibration regime.** The 95% $\mathrm{CrI}^{(r)}_{0.95}$ credible interval's coverage $\hat{C}$ (and its 95% confidence interval) of the AIPW generalized posterior with an uncalibrated $\omega = 1$, the plug-in $\hat{\omega}$, and the bootstrap $\hat{\omega}$ (Table 6) across 9 synthetic datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_9\}$. Results are computed from $R = 500$ repetitions at sample size $n = 1000$. Closest to exact frequentist calibration at 95% is in **bold**.

D.4 Comparison with a nonparametric Bayesian baseline

We compare the AIPW-based generalized posterior with a Bayesian Additive Regression Trees (BART) response-surface model, implemented using PyMC-BART, on $\mathcal{D}_1$ and $\mathcal{D}_9$. Both methods are evaluated on the same $R = 50$ simulated datasets with $n = 500$ observations.

DGP	Method	Coverage	CrI length	MSE	Time [s]
$\mathcal{D}_1$	BART (response surface)	0.980 (0.894-0.999)	0.745	0.0321	33.54
	Generalized Bayes (AIPW, ours)	0.980 (0.894-0.999)	0.499	0.0128	0.04
$\mathcal{D}_9$	BART (response surface)	1.000 (0.929-1.000)	2.195	0.2949	30.51
	Generalized Bayes (AIPW, ours)	0.960 (0.863-0.995)	0.777	0.0275	0.04

Table 8: **Comparison to BART.** The 95% credible interval's coverage $\hat{C}$ (and its 95% confidence interval), mean CrI length, MSE of the posterior mean, and mean runtime per repetition for an off-the-shelf BART response-surface model (PyMC-BART, 50 trees, one chain) and the AIPW generalized posterior, fitted to the same datasets. Results are computed from $R = 50$ paired repetitions at sample size $n = 500$; at this R, coverage differences are not resolved.

Table 8 shows that, while both approaches achieve high empirical coverage, our generalized posterior provides substantially shorter credible intervals and lower estimation error at considerably lower computational cost. Given the limited number of repetitions, we do not draw conclusions from small differences in coverage.

D.5 Sensitivity to cross-fitting

We investigate the effect of the number of cross-fitting folds $K \in \{2, 5, 10\}$ on the AIPW-based generalized posterior, considering $\mathcal{D}_6$ and $\mathcal{D}_8$ with $n = 1000$ and $R = 500$ repetitions. To isolate the effect of K, we use the same calibration parameter $\hat{\omega}$ across all configurations.

Table 9 shows that $K = 5$ and $K = 10$ yield similar performance, whereas using only two folds can substantially degrade coverage. This supports $K = 5$ as a reasonable practical default, although increasing the number of folds does not resolve the difficulties caused by limited overlap. Naturally, this recommendation is sensitive to sample size and choice of nuisance estimation.

	$\mathcal{D}_6$		$\mathcal{D}_8$	
Folds	Coverage	MSE	Coverage	MSE
$K = 2$	0.560 (0.515-0.604)	0.129	0.778 (0.739-0.814)	0.072
$K = 5$	0.540 (0.495-0.584)	0.130	0.964 (0.944-0.979)	0.029
$K = 10$	0.554 (0.509-0.598)	0.129	0.972 (0.953-0.985)	0.025

Table 9: **Choice of cross-fitting folds K.** The 95% credible interval's coverage $\hat{C}$ (and its 95% confidence interval) and MSE of the posterior mean of the AIPW generalized posterior for $K \in \{2, 5, 10\}$ on $\mathcal{D}_6$ (limited overlap) and $\mathcal{D}_8$ (high-dimensional sparse confounding). Results are computed from $R = 500$ repetitions at sample size $n = 1000$. All arms share the same $\hat{\omega}$, so the mean CrI length does not change with K (0.607 on $\mathcal{D}_6$, 0.668 on $\mathcal{D}_8$).

D.6 Sensitivity to nuisance estimation

We investigate the effect of different nuisance estimators on the AIPW-based generalized posterior, considering linear/ridge regression, tree ensembles, neural networks, and Bayesian Gaussian process posterior-mean plug-ins. Tables 10 and 11 show that faithful uncertainty quantification can be achieved with both frequentist and Bayesian nuisance estimators, but depends on their estimation accuracy, consistent with our theoretical rate conditions. A broader comparison across all nine settings is provided in Table 12.

		$\mathcal{D}_4$		$\mathcal{D}_5$		$\mathcal{D}_9$	
Nuisance learner	Bayesian	Coverage	MSE	Coverage	MSE	Coverage	MSE
Linear/ridge	✗	**0.952** (0.929-0.969)	0.006	**0.934** (0.909-0.954)	0.008	**0.958** (0.937-0.974)	0.013
Tree ensemble	✗	0.892 (0.861-0.918)	0.010	0.822 (0.786-0.855)	0.015	0.980 (0.964-0.990)	0.011
Small MLP	✗	0.882 (0.850-0.909)	0.010	0.280 (0.241-0.322)	0.051	0.900 (0.870-0.925)	0.023
Bayesian GP (RFF) post. mean	✓	0.956 (0.934-0.972)	0.006	0.804 (0.766-0.838)	0.017	0.992 (0.980-0.998)	0.010

Table 10: **Flexible and Bayesian nuisance estimators.** The 95% credible interval's coverage $\hat{C}$ (and its 95% confidence interval) and MSE of the posterior mean of the AIPW generalized posterior with four nuisance learners, on $\mathcal{D}_4$, $\mathcal{D}_5$, and $\mathcal{D}_9$. The Bayesian learner plugs in the fold-wise posterior mean of a Gaussian process (random-feature approximation). $\hat{\omega}$ is shared across learners, so the CrI length is identical within a dataset. Results are computed from $R = 500$ repetitions at sample size $n = 1000$ with $K = 5$ folds. Closest to exact frequentist calibration at 95% is in **bold**.

DGP	Cov.: GP post. mean	Cov.: linear/ridge	MSE: GP	MSE: ridge
$\mathcal{D}_4$	0.956 (0.934-0.972)	0.952 (0.929-0.969)	0.006	0.006
$\mathcal{D}_5$	0.804 (0.766-0.838)	0.934 (0.909-0.954)	0.017	0.008
$\mathcal{D}_9$	0.992 (0.980-0.998)	0.958 (0.937-0.974)	0.010	0.013

Table 11: **Bayesian nuisance posterior mean as plug-in.** The 95% credible interval's coverage $\hat{C}$ (and its 95% confidence interval) and MSE of the posterior mean when the nuisances are the fold-wise posterior mean of a Gaussian process (random-feature approximation), compared with the linear/ridge baseline. Results are computed from $R = 500$ repetitions at sample size $n = 1000$ with $K = 5$ folds and a shared $\hat{\omega}$. Coverage is near nominal on $\mathcal{D}_4$ and $\mathcal{D}_9$ but not on $\mathcal{D}_5$, consistent with the nuisance-rate condition.

Nuisance learner	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$
Linear/ridge	0.980 (0.894-0.999)	0.920 (0.808-0.978)	0.980 (0.894-0.999)	**0.940** (0.835-0.987)	**0.940** (0.835-0.987)	0.620 (0.472-0.753)	**0.960** (0.863-0.995)	**1.000** (0.929-1.000)	1.000 (0.929-1.000)
Tree ensemble	**0.920** (0.808-0.978)	**0.940** (0.835-0.987)	**0.940** (0.835-0.987)	0.860 (0.733-0.942)	0.860 (0.733-0.942)	0.400 (0.264-0.548)	0.920 (0.808-0.978)	1.000 (0.929-1.000)	**0.960** (0.863-0.995)
Small MLP	1.000 (0.929-1.000)	0.900 (0.782-0.967)	0.980 (0.894-0.999)	0.920 (0.808-0.978)	0.260 (0.146-0.403)	**0.740** (0.597-0.854)	0.960 (0.863-0.995)	1.000 (0.929-1.000)	0.920 (0.808-0.978)
Bayesian GP (RFF) post. mean	0.980 (0.894-0.999)	0.940 (0.835-0.987)	1.000 (0.929-1.000)	0.920 (0.808-0.978)	0.800 (0.663-0.900)	0.360 (0.229-0.508)	0.940 (0.835-0.987)	0.860 (0.733-0.942)	1.000 (0.929-1.000)

Table 12: **Nuisance learners across all datasets.** The 95% $\mathrm{CrI}^{(r)}_{0.95}$ credible interval's coverage $\hat{C}$ (and its 95% confidence interval) of the AIPW generalized posterior for each nuisance learner across 9 synthetic datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_9\}$. Results are computed from $R = 50$ repetitions at sample size $n = 1000$ with $K = 5$ folds and a shared $\hat{\omega}$. Closest to exact frequentist calibration at 95% is in **bold**.

D.7 Semi-synthetic IHDP benchmark

We further evaluate ATE estimation on the semi-synthetic IHDP benchmark, considering both IHDP-100 and IHDP-1000 replications. Table 13 presents the results for the RA, IPW, and AIPW-based generalized posteriors. Unlike in most of our synthetic settings, the AIPW-based posterior does not achieve nominal coverage on this benchmark, illustrating that Neyman orthogonality alone does not guarantee faithful uncertainty quantification in finite samples.

		IHDP-100 ($R = 100$)			IHDP-1000 ($R = 1000$)		
Strategy	Orthogonal	Coverage	CrI length	MSE	Coverage	CrI length	MSE
RA	✗	0.420 (0.322-0.523)	0.862	2.098	0.392 (0.362-0.423)	0.883	1.808
IPW	✗	0.990 (0.946-1.000)	6.715	2.251	0.988 (0.979-0.994)	6.860	4.030
AIPW	✓	0.730 (0.632-0.814)	1.286	2.212	0.688 (0.658-0.717)	1.307	1.786

Table 13: **Semi-synthetic IHDP benchmark (ATE).** The 95% credible interval's coverage $\hat{C}$ (and its 95% confidence interval), mean CrI length, and MSE of the posterior mean for the generalized posteriors of each strategy $S \in \{\mathrm{RA}, \mathrm{IPW}, \mathrm{AIPW}\}$ on the IHDP-100 ($R = 100$) and IHDP-1000 ($R = 1000$) replications under a diffuse prior $\mathcal{N}(0, 100^2)$. On IHDP, the AIPW posterior under-covers.

E Experiment Details

E.1 Implementation Details

The code for all experiments is available at `https://github.com/EmilJavurek/Generalised-Bayes-for-Causal-Inference`. Here, we provide a short description of the essential implementation details.

Nuisances: For the nine-setting ATE benchmark and the baseline configurations of our sensitivity experiments, we fit logistic regression for the propensity and two ridge regressions for the outcomes (T-learner). These models use a shared, fixed feature map $\Phi(X) = [X, X^2, \sin(X), x_{\text{extra}}, 1]$, where $x_{\text{extra}} = X_1 X_2$ if $d \geq 2$ (or X_1^3 if $d = 1$). Propensities are regularized with λ_{prop} and clipped to $[\epsilon, 1 - \epsilon]$ to stabilize IPW/AIPW. We use cross-fitting: nuisance models are fit on each training fold and pseudo-outcomes are evaluated on held-out folds. The nuisance-estimation sensitivity experiments additionally consider tree ensembles, neural networks, and Bayesian Gaussian process posterior-mean plug-ins.

Prior: For the nine-setting ATE benchmark, we use a Gaussian prior $\mathcal{N}(\mu_0, \sigma_0^2)$ with $\mu_0 = 0$ and $\sigma_0 = 1$. The additional ATE robustness and sensitivity experiments instead use a diffuse $\mathcal{N}(0, 100^2)$ prior, unless otherwise specified. In the prior-strength experiment, we explicitly vary both the prior mean and variance. For CATE, the prior is a Gaussian process with the kernel specified in the configuration; we mainly use RBF or Matern kernels.

Variational family: For the nine-setting ATE benchmark, the variational family is a univariate Gaussian with parameters $(\mu, \log \sigma)$; optimizing in $\log \sigma$ enforces positivity of σ and yields stable optimization. When optimization is stopped early, the learned posterior variance can be slightly conservative relative to the analytic closed form. For CATE, the variational family is a sparse inducing-point GP with M inducing locations initialized from the training covariates. In the example CATE configuration, we use $M = 20$ and a Matern-$5/2$ kernel with lengthscale 2.0, variance 2.0, and jitter 10^{-4}. A constant mean function is fit to the pseudo-outcomes and subtracted before GP fitting, and then added back to predictions.

Variational inference (VI): We use the Adam optimizer. For the nine-setting ATE benchmark, we use a batch size of 200 (number of θ samples per update), learning rate 0.03, and 2000 epochs. For CATE, the batch size field is ignored because the GP loss is evaluated in full batch.

Additional ATE experiments: For the misspecification, calibration, prior-strength, cross-fitting, nuisance-estimation, and semi-synthetic evaluations, we compute the generalized posterior analytically using Gaussian conjugacy for squared-error pseudo-outcome losses, rather than VI. Unless otherwise specified, we use $K = 5$ cross-fitting folds and the analytic plug-in calibration $\hat{\omega} = \widehat{\mathrm{Var}}(\hat{Y}^S)^{-1}$. The calibration comparison also considers a bootstrap estimate obtained from 100 resamples of the fixed cross-fitted pseudo-outcomes, without refitting nuisance models. This differs from the coverage-matching bootstrap calibration discussed in Section 5.3. In the cross-fitting and nuisance-learner ablations, we hold $\hat{\omega}$ fixed at the value obtained from the $K = 5$ linear/ridge baseline on the same dataset, isolating the effects of the respective choices. The Bayesian nuisance experiment uses fold-wise posterior-mean plug-ins without integrating over nuisance posterior uncertainty.

E.2 Data-Generating Processes

E.2.1 Back-door adjustment setting

1. $\mathcal{D}_1$ *(Linear, confounded, homoskedastic).*

$$\begin{aligned} X &\sim \mathcal{N}(0, I_2), \\ e(X) &= \sigma(X^\top \beta), \qquad \sigma(u) = \frac{1}{1+e^{-u}}, \\ A \mid X &\sim \text{Bernoulli}\big(e(X)\big), \\ Y \mid (A, X) &= \tau\, A + X^\top \gamma + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1). \end{aligned}$$

   The CATE is $\tau(X) = \tau$ and the ATE is $\text{ATE} = \tau$.

2. $\mathcal{D}_2$ *(Linear CATE heterogeneity, mean-zero covariates).*

$$\begin{aligned} X &\sim \mathcal{N}(0, I_2), \\ e(X) &= \sigma(X^\top \beta), \\ A \mid X &\sim \text{Bernoulli}\big(e(X)\big), \\ Y \mid (A, X) &= \big(\theta_0 + X^\top \theta\big) A + X^\top \gamma + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1). \end{aligned}$$

   The CATE is $\tau(X) = \theta_0 + X^\top \theta$ and the ATE is $\text{ATE} = \theta_0$.

3. $\mathcal{D}_3$ *(Linear CATE heterogeneity, nonzero mean covariates).*

$$\begin{aligned} X &\sim \mathcal{N}(\mu, I_2), \\ e(X) &= \sigma(X^\top \beta), \\ A \mid X &\sim \text{Bernoulli}\big(e(X)\big), \\ Y \mid (A, X) &= \big(\theta_0 + X^\top \theta\big) A + X^\top \gamma + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1). \end{aligned}$$

   The CATE is $\tau(X) = \theta_0 + X^\top \theta$ and the ATE is $\text{ATE} = \theta_0 + \theta^\top \mu$.

4. $\mathcal{D}_4$ *(Nonlinear outcome; quadratic and interaction CATE).*

$$\begin{aligned} X &\sim \mathcal{N}(0, I_2), \\ e(X) &= \sigma(X^\top \beta), \\ A \mid X &\sim \text{Bernoulli}\big(e(X)\big), \\ f(X) &= X_1^2 + \sin(X_2), \\ Y \mid (A, X) &= \big(\alpha_0 + \alpha_1 X_1 X_2\big) A + f(X) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1). \end{aligned}$$

   The CATE is $\tau(X) = \alpha_0 + \alpha_1 X_1 X_2$ and the ATE is $\text{ATE} = \alpha_0$.

5. $\mathcal{D}_5$ *(Nonlinear propensity; linear outcome).*

$$\begin{aligned} X &\sim \mathcal{N}(0, I_2), \\ z(X) &= b_0 + b_1 X_1 + b_2 X_1^2 + b_3 \sin(X_2), \\ e(X) &= \sigma\big(z(X)\big), \\ A \mid X &\sim \text{Bernoulli}\big(e(X)\big), \\ h(X) &= X_1 + \tfrac{1}{2} X_1^2 + \tfrac{1}{2} \sin(X_2), \\ Y \mid (A, X) &= \tau\, A + h(X) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1). \end{aligned}$$

   The CATE is $\tau(X) = \tau$ and the ATE is $\text{ATE} = \tau$.

6. $\mathcal{D}_6$ *(Limited overlap).*

$$\begin{aligned} X &\sim \mathcal{N}(0, I_2), \\ e(X) &= \sigma(3.5 + 3X_1), \\ A \mid X &\sim \text{Bernoulli}\big(e(X)\big), \\ Y \mid (A, X) &= \tau\, A + X^\top \gamma + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1). \end{aligned}$$

   The CATE is $\tau(X) = \tau$ and the ATE is $\text{ATE} = \tau$.

7. $\mathcal{D}_7$ *(Heteroskedastic heavy-tailed noise).*

$$\begin{aligned} X &\sim \mathcal{N}(0, I_2), \\ e(X) &= \sigma(X^\top \beta), \\ A \mid X &\sim \text{Bernoulli}\big(e(X)\big), \\ \sigma(X) &= \exp\big(0.5 X_1\big), \\ \eta &\sim t_\nu, \\ Y \mid (A, X) &= \tau\, A + X^\top \gamma + \sigma(X)\, \eta. \end{aligned}$$

The CATE is $\tau(X) = \tau$ and the ATE is $\text{ATE} = \tau$.

8. $\mathcal{D}_8$ *(High-dimensional sparse confounding).*

$$\begin{aligned} X &\sim \mathcal{N}(0, I_p), \\ \beta_j = 0,\ \gamma_j &= 0 \quad \text{for } j > s, \\ \beta_{1:s} = \text{linspace}(0.2, 1.0, s), &\qquad \gamma_{1:s} = \text{linspace}(1.0, 0.2, s), \\ e(X) &= \sigma(X^\top \beta), \\ A \mid X &\sim \text{Bernoulli}\big(e(X)\big), \\ Y \mid (A, X) &= \tau\, A + X^\top \gamma + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1). \end{aligned}$$

The CATE is $\tau(X) = \tau$ and the ATE is $\text{ATE} = \tau$.

9. $\mathcal{D}_9$ *(Friedman-style nonlinear CATE).*

$$\begin{aligned} X &\sim \text{Unif}\big([0,1]^5\big), \\ e(X) &= \sigma(-0.5 + X_1 - 0.25 X_2 + 0.25 X_3)\,, \\ A \mid X &\sim \text{Bernoulli}\big(e(X)\big), \\ \mu(X) &= 10 \sin(\pi X_1 X_2) + 20 (X_3 - 0.5)^2 + 10 X_4 + 5 X_5, \\ \tau(X) &= 1 + \frac{X_1}{X_2 + 0.1}, \\ Y \mid (A, X) &= \mu(X) + \tau(X) A + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1). \end{aligned}$$

The CATE is $\tau(X) = 1 + \frac{X_1}{X_2 + 0.1}$ and the ATE is $\text{ATE} = 1 + \frac{1}{2}\log(11)$.

F Limitations

Our framework inherits the assumptions required for causal identification and for orthogonal statistical learning. In the back-door ATE and CATE settings considered here, violations of consistency, positivity, unconfoundedness, no interference, or accurate measurement can lead the generalized posterior to quantify uncertainty around an identified statistical functional that differs from the causal estimand of interest. The theoretical guarantees further rely on regularity conditions such as Bernstein–von Mises approximations, Neyman-orthogonality, cross-fitting, and sufficient nuisance-estimation rates; these may fail under weak identification, unstable nuisance learning, highly irregular losses, small samples, or very high-dimensional targets. Practical performance can also depend on implementation choices such as nuisance learners, variational approximations, Gaussian process kernels, inducing points, and optimizer settings.

Broader Impact

This work develops a methodological framework for uncertainty quantification in causal inference. A potential positive impact is that calibrated uncertainty estimates can help practitioners avoid overconfident causal conclusions, especially in domains where treatment-effect estimates may inform consequential decisions such as medicine, public policy, education, or economics. By placing priors directly on causal estimands and using identification-driven losses, the framework may also make uncertainty statements more transparent and easier to scrutinize than approaches that encode beliefs indirectly through high-dimensional nuisance models.

At the same time, causal uncertainty estimates are only meaningful under the assumptions required for identification and estimation. If the relevant causal assumptions fail, for example due to unmeasured confounding, violations of positivity, interference, measurement error, or distribution shift, then calibrated-looking credible intervals may still support incorrect causal conclusions. This risk is especially important in high-stakes applications, where erroneous treatment-effect estimates could lead to harmful or unfair decisions. Moreover, because the proposed framework can be combined with flexible machine learning nuisance estimators, it may inherit biases present in the data or in the chosen learning pipeline, including biases that affect subpopulations differently.

We therefore view the proposed framework as a tool for uncertainty-aware causal analysis rather than as a substitute for domain expertise, study design, or sensitivity analysis. In practical deployments, users should explicitly state the causal identification assumptions, assess overlap and nuisance-model quality, report calibration and sensitivity diagnostics, and avoid using the resulting posterior summaries as standalone justification for high-stakes interventions. The experiments in this paper comprise synthetic simulations and a semi-synthetic IHDP benchmark; we do not release human-subject datasets, pretrained generative models, or other high-risk assets.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the proposed generalized Bayesian framework, its scope across ATE/CATE and loss-based causal estimators, the Neyman-orthogonal robustness guarantee, and the empirical claim of calibrated uncertainty. These claims are developed in Sections 5–6 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix F discusses limitations arising from causal identification assumptions, regularity conditions for the theoretical guarantees, nuisance estimation, calibration, and practical implementation choices.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the assumptions for the main posterior-stability result and provides formal finite-dimensional and projected infinite-dimensional versions with proofs in Appendix A. The informal theorem in Section 5.4 is supported by the formal statements, assumptions, and proof details in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 6 and Appendices D and E describe the experimental protocols, evaluation metrics, synthetic data-generating processes, semi-synthetic IHDP benchmark, nuisance estimation, calibration, and posterior computation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides a link to the experiment code repository. Appendix E documents the implementation and synthetic data-generating processes, while Appendix D reports additional experiments, including the semi-synthetic IHDP evaluation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 and Appendix E specify the experimental settings, sample sizes, repetitions, nuisance estimators, cross-fitting, priors, posterior computation, optimization settings, and calibration procedures. Additional experimental configurations are described in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports empirical coverage with 95% confidence intervals and reports credible-interval lengths with standard deviations across repetitions. The variability is computed over repeated runs with different random seeds/synthetic samples.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments have been run on a laptop `CPU` with trivial runtime (<1hr).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is methodological and is evaluated using synthetic simulations and a semi-synthetic benchmark based on existing data. No new human-subject data are collected, and no deployed decision systems are involved. To the best of the authors' knowledge, the research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Broader Impact section discusses potential benefits of calibrated causal uncertainty quantification and risks associated with identification failures, model misspecification, biased nuisance estimation, and the misuse of causal conclusions in consequential decisions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, or other assets with a high risk of misuse. The released assets consist of code for methodological evaluation using synthetic and semi-synthetic experiments.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The experiments use synthetic data-generating processes rather than existing datasets, pretrained models, or third-party data assets. Standard software dependencies, if any, should be documented in the released code repository.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper releases code for the proposed generalized Bayes causal inference experiments. The accompanying appendix documents the experimental setup, synthetic data-generating processes, nuisance estimators, priors, and optimization details.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The research does not recruit participants, conduct crowdsourcing, or collect new human-subject data. The semi-synthetic IHDP evaluation uses an existing benchmark with simulated outcomes.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work involves no participant recruitment or new human-subject data collection. The experiments use synthetic data and an existing semi-synthetic benchmark; no new human-subject study was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development, theoretical results, and experiments do not involve LLMs as an important, original, or non-standard component. Any LLM use solely for writing, editing, or formatting would not require declaration under the checklist question.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.