A Causal Target for Learning to Defer Under Hidden Confounding

Yanmin Li[1] *, **Lihua Liu**[1] *, **Xin Wang**[2], **Zhilong Mao**[1], **Jibing Wu**[1] †, **Weidong Bao**[1],

[1]Laboratory for Big Data and Decision, National University of Defense Technology
[2]University of Science and Technology of China
yanminli@nudt.edu.cn, lihualiu@nudt.edu.cn, wz520@mail.ustc.edu.cn, {mzl02, wujibing, wdbao}@nudt.edu.cn

Abstract

Learning decision policies from confounded observational data is a challenging task in causal inference, as unobserved confounders can lead to biased or suboptimal actions when relying solely on machine learning models. A synergistic approach is learning to defer, which decides when to act itself and when to defer to a human expert with access to unobserved information. However, constructing the learning target, which defines the probability of choosing each action or deferral, remains a core challenge. To address this, we propose causal-target-based learning to defer (CTLD) framework, where the causal target is constructed from sharp bounds on potential outcomes. Specifically, the degree of overlap between these bounds determines the probability of deferral, while their relative positions and widths define the probabilities over actions. CTLD aligns model predictions with this causal target to make probabilistic decisions over actions and deferral. We present comprehensive theoretical guarantees for the learned policy and demonstrate the effectiveness of CTLD on synthetic and semi-synthetic datasets.

Code — https://github.com/JustinLiam/CTLD

Introduction

Causal policy learning, which aims to optimize decision-making from observational data, is critical in high-stakes domains such as healthcare, finance, and social services (Kallus 2018; Athey and Wager 2021). However, a fundamental challenge in leveraging such data is hidden confounding, which induces structural bias in causal effect estimation, thereby compromising the validity of learned policies and leading to suboptimal or even detrimental decision-making (Imbens and Rubin 2015). For example, in healthcare, unrecorded lifestyle or psychological factors may influence both treatment choices and outcomes, thereby serving as unobserved confounders that distort the data-driven policy learning.

Recent studies on confounding-robust policy learning, such as those based on the minimax regret framework (Kallus and Zhou 2018, 2021), often adopt a worst-case perspective, resulting in overly conservative treatments with limited utility. To address this limitation, we draw inspiration from the emerging paradigm of "Learning to Defer" (L2D) (Mozannar and Sontag 2020), where a model is trained not only to act autonomously but also to defer to a human expert when facing high causal uncertainty. This strategy may alleviate the pitfalls of overly conservative policies, thus potentially achieving a better balance between robustness and effectiveness, as illustrated in Figure 1.

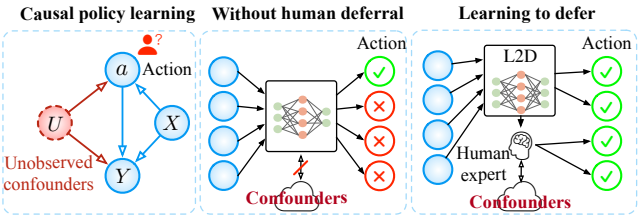


Figure 1: Overview of causal policy learning problem setup and two learning paradigms. Policies without deferral may yield suboptimal actions. The learning to defer (L2D) enables the model to defer to a human expert, thereby mitigating confounding effects and making better action decisions.

The L2D paradigm can be viewed as learning a classifier with the option of deferring to an expert. However, the observational data only reflects the covariates, action (e.g., treatment 0 or 1) and corresponding outcome, without explicit labels for when to defer. Thus, rather than learning to imitate deferrals, the model must infer when to defer by quantifying its uncertainty about the best course of action under hidden confounding. The core challenge, therefore, is to construct a learning target that serves as a principled proxy for the unknowable optimal decision, guiding the model on when to act versus when to defer.

Prior work addresses this by designing a cost-sensitive target which measures the penalty among actions and deferral (Ghoummaid and Shalit 2024). A key weakness of this approach is that its deferral cost directly incorporates the raw observed outcome for each sample, which makes the target sensitive to outcome stochasticity and may limit the learned policy's reliability. Furthermore, this framework triggers deferral whenever the deferral penalty is lower than the penalties for taking actions. This trigger is based on a

*These authors contributed equally.
†Jibing Wu is the corresponding author.

simple cost comparison, rather than a direct quantification of the causal uncertainty between actions. These limitations highlight a clear research gap in designing a robust learning target based on a direct quantification of causal uncertainty.

To address this gap, we propose the causal-target-based learning to defer (CTLD) framework to learn deferral policies. In the presence of hidden confounding, the true causal effects of actions are not identifiable, but they can be tightly bounded by sharp bounds[1]. By leveraging these bounds, CTLD constructs a probabilistic learning target designed to capture the uncertainty under hidden confounding. Specifically, the deferral probability is determined by the degree of overlap between the bounds, while the action probabilities are determined by the estimated treatment effect, which is in turn tempered by the overall uncertainty captured by the total width of the bounds. A model is then trained to match this target using a tailored loss function and an asymmetric softmax parameterization, which is suitable for the distinct nature of action versus deferral decisions, thereby learning a mapping from covariates to a probability distribution over actions and deferral.

Our main contributions are as follows:

- We introduce a CTLD framework that leverages sharp causal bounds to construct the stable, probabilistic causal target, enabling robust action and deferral decisions under hidden confounding.
- We establish theoretical guarantees for CTLD, including a regret-transfer bound that links the training loss to the performance gap between the learned policy and the optimal policy, while a generalization bound ensuring that the learned policy remains reliable on unseen data.
- We conduct extensive experiments on both synthetic and semi-synthetic datasets, demonstrating that CTLD outperforms existing baselines in both policy values and deferral behavior.

Related Work

Causal Policy Learning. Learning decision policies from observational data is fundamentally a causal inference challenge (Imbens and Rubin 2015). A core difficulty is that unobserved confounders, which are factors influencing both past actions and outcomes, can create spurious correlations, making it difficult to discern which actions are genuinely effective. To directly confront this problem, the field of causal policy learning seeks to find not just optimal, but "robust" policies, whose performance is guaranteed to be safe even under worst-case assumptions about the confounding.

Broadly, two main approaches have emerged. Early "weight-centric" method sought robustness through data reweighting, exemplified by minimax regret frameworks using the Inverse Propensity Weighting (IPW) estimator (Kallus and Zhou 2018, 2021). The statistical instability of this strategy, particularly its tendency to assign extremely large importance to a few rare observations, spurred the evolution towards a more direct, "value-centric" perspective. This latter approach bypasses data re-weighting and instead asks, "Given the potential for confounding, what are the best-case and worst-case outcomes for any given action?", leading to more stable methods that derive sharp bounds directly on potential outcomes (Dorn and Guo 2023; Hess et al. 2025).

Learning to Defer Paradigm. A promising avenue for mitigating risk under uncertainty is to design systems that can defer to a human expert (Bondi et al. 2022; Hemmer et al. 2023; Ruggieri and Pugnana 2025). This principle, known as "learning to defer" (L2D), has been systematically established in supervised learning, allowing a model to abstain on difficult instances and pass them to an expert, thereby improving overall reliability (Mozannar et al. 2023; Tailor et al. 2024). The established approach for training such systems reduces the L2D problem to a well-understood, cost-sensitive classification task using consistent surrogate losses (Mozannar and Sontag 2020). These developments provides a strong, principled foundation for building systems that can intelligently manage the trade-off between autonomous action and expert referral.

However, applying the L2D to causal policy learning introduces a fundamental challenge not present in supervised learning: the absence of ground-truth labels for counterfactual outcomes. To overcome this lack of direct supervision, all methods must rely on a set of assumptions to estimate causal effects from observational data. To make the problem tractable, pioneering works often operate under the assumption of unconfoundedness (e.g., Leitão et al. (2022); Gao et al. (2021)). This has led to the development of simplified systems that learn to route decisions between an AI and an expert, optimizing the team's overall performance by leveraging their complementary strengths. This unconfoundedness assumption, while powerful, is untenable in many real-world applications, as unobserved confounders can lead to biased estimates and, consequently, unsafe policies.

Learning to Defer under Hidden Confounding. Recently, a few pioneering works have begun to tackle the challenging problem of learning a deferral policy in the presence of hidden confounding. These efforts can be broadly categorized into three main methodologies. The first is **direct interval estimation** of the conditioal average treatment effect (CATE), an approach taken by (Jesson et al. 2021), who first estimate a range of possible CATE values and then apply a simple rule, such as deferring if the estimated interval contains zero. A second methodology is the **"weight-centric" approach**, pioneered by (Gao and Yin 2025), who propose a minimax framework that optimizes a worst-case policy risk over an uncertainty set of importance weights. However, this method can suffer from high variance due to the instability of these weights. A third strategy is **cost-sensitive classification**, as developed by (Ghoummaid and Shalit 2024), whose training target is prone to instability from its reliance on noisy outcomes and provides only an indirect proxy for causal uncertainty. These limitations across different methodologies highlight a clear research gap in designing a robust learning target based on a direct quantification of causal uncertainty. A structured comparison of these

[1]Sharp bounds refer to the tightest possible upper and lower limits on potential outcomes that are consistent with the observed data and the assumed level of hidden confounding (Ho and Rosen 2015).

approaches is provided in Appendix A.

Our Method

Problem Setup

We consider a dataset $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^N$, where for each unit i, $X_i \in \mathcal{X}$ represents pre-treatment covariates, $A_i \in \mathcal{A}$ is the treatment received, and $Y_i \in \mathbb{R}$ is the observed outcome. In this work, we focus on the binary treatment setting, where the treatment action is $A_i \in \{0, 1\}$. Under the potential outcomes framework, $Y_i(0)$ and $Y_i(1)$ denote the two potential outcomes for unit i. A summary of key terms and notation is included in Appendix B for reference.

The traditional goal of policy learning is to find a policy $\pi : \mathcal{X} \to \{0, 1\}$ that maximizes the expected outcome, $V(\pi) = \mathbb{E}[Y(\pi(X))]$. Reliably evaluating this value from observational data, depends on three core assumptions: the stable unit treatment value assumption (SUTVA), positivity, and unconfoundedness (Imbens and Rubin 2015). However, the Unconfoundedness assumption (i.e., $\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$) is often the most challenging to satisfy in practice (Gao and Yin 2025; Hess et al. 2025).

The violation of unconfoundedness can lead to suboptimal and unreliable policies. To mitigate this risk, we adapt the L2D framework, a form of human-AI collaboration. The goal is to learn a deferral policy $\pi : \mathcal{X} \to \{0, 1, \perp\}$, which maps covariates either to a specific treatment action or to the deferral action ($\perp$). The value of such a policy π is the expected outcome of the joint system:

$$V(\pi) = \mathbb{E}\Big[\mathbb{I}(\pi(X) \neq \perp) Y(\pi(X)) + \mathbb{I}(\pi(X) = \perp) Y(A)\Big] \tag{1}$$

where the system follows the AI's decision $\pi(X)$ if it does not defer, and the expert's decision otherwise. Following Ghoummaid and Shalit (2024), we assume the historical data was generated by this same expert, and model the expert's action upon deferral as the historically observed action.

Our central problem is to learn an optimal policy $\pi^* = \arg\max_\pi V(\pi)$ in the presence of hidden confounding. To formalize this, we relax the unconfoundedness assumption by assuming the existence of an unobserved confounder U such that unconfoundedness holds conditioned on both X and U:

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A|X, U \tag{2}$$

To make this assumption operational, we quantify the strength of the confounding induced by U using the marginal sensitivity model (MSM) (Tan 2006), which bounds the degree to which U can influence treatment assignment given the estimable nominal propensity score $e(x) := P(A = 1|X = x)$ versus the unidentifiable true propensity score $e(x, u) := P(A = 1|X = x, U = u)$.

Assumption 1 (MSM Assumption)**.** *The ratio between the true treatment odds, $e(x, u)/(1 - e(x, u))$, and the nominal treatment odds, $e(x)/(1 - e(x))$, is almost surely bounded by a factor $\Lambda \geq 1$:*

$$\Lambda^{-1} \leq \frac{e(x, u)}{1 - e(x, u)} \Big/ \frac{e(x)}{1 - e(x)} \leq \Lambda.$$

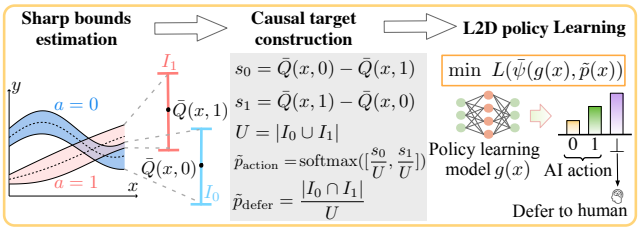


Figure 2: Overview of the CTLD framework.

A larger Λ accommodates greater degrees of unobserved confounding. When $\Lambda = 1$, the assumption reduces to standard unconfoundedness.

Causal-Target-based Learning to Defer

To learn a deferral policy under hidden confounding, our CTLD method proceeds in three key steps, as shown in Figure 2: first, we estimate sharp bounds on potential outcomes to quantify causal uncertainty; second, we leverage these bounds to construct a novel causal target; and finally, we train a policy to match this target using a tailored optimization procedure.

Sharp Bounds Estimation The first step of our CTLD framework is to quantify the causal uncertainty arising from hidden confounders. We achieve this by estimating the sharp upper and lower bounds, denoted $Q^\pm(x, a)$, for the conditional average potential outcome (CAPO), where $Q(x, a) = \mathbb{E}[Y(a)|X = x]$.

These bounds are derived under Assumption 1 and are formally presented in Definition 1, which builds upon the quantile balancing method of Dorn and Guo (2023). The full derivation is provided in Appendix C.

Definition 1 (Sharp Bounds for CAPO)**.** *Under Assumption 1, the sharp upper and lower bounds on the CAPO $Q(x, a)$, denoted $Q^\pm(x, a)$, are given by:*

$$Q^\pm(x, a) = w^\pm(x, a)\bar{\mu}^\pm(x, a) + w^\mp(x, a)\mu^\pm(x, a) \tag{3}$$

where the components are defined as follows. The weighting coefficients $w^\pm(x, a)$ are:

$$w^\pm(x, 1) = e(x) + (1 - e(x))\Lambda^{\pm 1},$$
$$w^\pm(x, 0) = (1 - e(x)) + e(x)\Lambda^{\pm 1}.$$

The partial expectations are:

$$\mu^\pm(x, a) = \mathbb{E}[Y \cdot \mathbb{I}^\pm(Y, x, a)|X = x, A = a],$$
$$\bar{\mu}^\pm(x, a) = \mathbb{E}[Y \cdot \bar{\mathbb{I}}^\pm(Y, x, a)|X = x, A = a].$$

The indicator functions $\mathbb{I}^\pm$ and $\bar{\mathbb{I}}^\pm$ are given by:

$$\mathbb{I}^+(y, x, a) = \mathbb{I}\{y \leq q_\tau(x, a)\},$$
$$\bar{\mathbb{I}}^+(y, x, a) = \mathbb{I}\{y > q_\tau(x, a)\},$$
$$\mathbb{I}^-(y, x, a) = \mathbb{I}\{y \leq q_{1-\tau}(x, a)\},$$
$$\bar{\mathbb{I}}^-(y, x, a) = \mathbb{I}\{y > q_{1-\tau}(x, a)\},$$

where the quantile level is $\tau = \Lambda/(1 + \Lambda)$.

In practice, we compute these bounds by first using neural networks to estimate the required nuisance functions: the nominal propensity score $e(x)$ and the conditional quantile function $q_\tau(x, a)$. The final estimated bounds, denoted $\hat{Q}^{\pm}(x, a)$, are obtained by plugging these nuisance estimates into Equation (3).

Causal Target Construction With the uncertainty bounds $\hat{Q}^{\pm}(x, a)$ established, the subsequent challenge is to convert this information into a robust learning target. Prior work, such as CARED (Ghoummaid and Shalit 2024), has approached this by constructing a surrogate cost that directly incorporates the single-sample observed outcome y. While innovative, this reliance makes the learning target sensitive to outcome stochasticity, which can impair the learned policy's reliability. Our core idea, therefore, is to construct a causal target that derives its structure from the stable geometry of the uncertainty intervals, rather than from the noisy observed outcome. This target vector is thus designed with two key components: a deferral probability to represent the degree of causal uncertainty, and a calibrated action distribution to reflect the preferred treatment.

Deferral Probability. In our framework, the decision to defer is determined by the uncertainty about which action is optimal. We quantify this uncertainty by the degree of overlap between the estimated CAPO intervals for the two actions. Let $I_a = [\hat{Q}^-(x, a), \hat{Q}^+(x, a)]$. The deferral probability is the ratio of the intersection length to the union length of these intervals:

$$\tilde{p}_{\text{defer}}(x) = \frac{|I_0 \cap I_1|}{|I_0 \cup I_1|}, \tag{4}$$

This construction intuitively ensures that significant overlap between the bounds leads to a high probability of deferral, while well-separated bounds result in a probability approaching zero.

Action Probability. For the action probabilities, we aim to capture the estimated treatment effect while being calibrated for uncertainty. To derive this probabilities, we begin with a provisional treatment effect based on the interval's midpoint, $\bar{Q}(x, a) = (\hat{Q}^+(x, a) + \hat{Q}^-(x, a))/2$. This midpoint, serving as an initial uncalibrated measure, is then used to define a raw score for action 1 as $\text{score}_1(x) = \bar{Q}(x, 1) - \bar{Q}(x, 0)$, with the score for action 0 being its negative.

However, using these raw scores directly can lead to overconfident probabilities, particularly when the uncertainty bounds are wide. To address this, we introduce an **uncertainty-scaling** mechanism. We scale the scores by the total uncertainty, $U(x) = |I_0 \cup I_1|$, before applying the softmax function:

$$\tilde{p}_{\text{action}}(x) = \text{softmax}\left(\left[\frac{\text{score}_0(x)}{U(x)}, \frac{\text{score}_1(x)}{U(x)}\right]\right). \tag{5}$$

This scaling tempers the model's confidence when causal estimates are imprecise.

Finally, this action distribution $\tilde{p}_{\text{action}}(x)$ and the deferral probability $\tilde{p}_{\text{defer}}(x)$ are combined to form the complete causal target $\tilde{p}(x)$:

$$\tilde{p}(x) = \big(\tilde{p}_0(x), \tilde{p}_1(x); \tilde{p}_{\text{defer}}(x)\big). \tag{6}$$

The resulting vector serves as a rich learning target that encodes both a calibrated action distribution and the degree of causal uncertainty.

L2D Policy Learning The final step is to learn a scoring function $g(x) : \mathcal{X} \to \mathbb{R}^3$ that maps input features to scores for each choice $\{0, 1, \perp\}$. Because the deferral option is conceptually separate from the mutually exclusive actions $\{0, 1\}$, a standard softmax over all three outputs is flawed. It would create an undesirable coupling, forcing the deferral probability to decrease as action probability increases, thereby ignoring the overall decision uncertainty.

To handle this structure correctly, we adopt the *Asymmetric Softmax parameterization* ($\bar{\psi}$) from Cao et al. (2023). This function maps the raw scores $g(x)$ to a valid probability vector suitable for our problem.

Definition 2 (Asymmetric Softmax Parameterization)**.** *For a score vector $u = [u_0, u_1, u_\perp]$, the transformation $\bar{\psi}(u) = [\bar{\psi}_{\textit{action}}(u); \bar{\psi}_{\textit{defer}}(u)]$ is defined as:*

$$\bar{\psi}_{\textit{action}}(u) = \textit{softmax}([u_0, u_1]),$$

$$\bar{\psi}_{\textit{defer}}(u) = \frac{\exp(u_\perp)}{\sum_{j\in\{0,1,\perp\}} \exp(u_j) - \max_{j\in\{0,1\}} \exp(u_j)}.$$

This parameterization ensures that the action probabilities $\bar{\psi}_{\text{action}}(u) = (\bar{\psi}_0(u), \bar{\psi}_1(u))$ form a valid distribution (i.e., $\bar{\psi}_0(u) + \bar{\psi}_1(u) = 1$), while the deferral probability $\bar{\psi}_{\text{defer}}(u)$ is a separate, calibrated value between 0 and 1. Our learning objective is to minimize the discrepancy between the model's output probabilities $\bar{\psi}(g(X))$ and our constructed causal target $\tilde{p}(X)$. We achieve this by minimizing a surrogate risk $\hat{R}(g)$, which is a weighted sum of two cross-entropy losses:

$$\begin{aligned}\hat{R}(g) = \mathbb{E}_{X\sim\hat{P}_n}\Big[&\mathcal{L}_{\text{action}}(g(X), \tilde{p}_{\text{action}}(X)) \\ &+ \lambda \cdot \mathcal{L}_{\text{defer}}(g(X), \tilde{p}_{\text{defer}}(X))\Big]\end{aligned} \tag{7}$$

where λ is a balancing hyperparameter. The action loss $\mathcal{L}_{\text{action}}$ is the standard cross-entropy between the predicted and target action distributions. The deferral loss $\mathcal{L}_{\text{defer}}$ is the binary cross-entropy for the deferral decision:

$$\begin{aligned}\mathcal{L}_{\text{defer}} = -\Big[&\tilde{p}_{\text{defer}}(X) \log \bar{\psi}_{\text{defer}}(g(X)) \\ &+ (1 - \tilde{p}_{\text{defer}}(X)) \log(1 - \bar{\psi}_{\text{defer}}(g(X)))\Big]\end{aligned} \tag{8}$$

The optimal scoring function g^* is found by minimizing this empirical risk over a given model class $\mathcal{G}$:

$$g^* = \arg\min_{g\in\mathcal{G}} \hat{R}(g). \tag{9}$$

The final policy $\pi^*(x)$ is then derived from these optimal scores. Specifically, after applying the Asymmetric Softmax transformation, we select the choice with the highest resulting probability. This ensures the decision rule is consistent with the calibrated probabilities used during training:

$$\pi^*(x) = \arg\max_{j\in\{0,1,\perp\}} \bar{\psi}_j(g^*(x)). \tag{10}$$

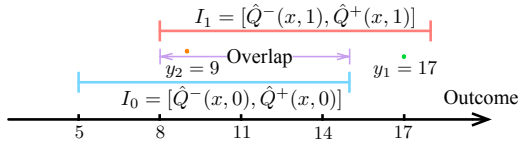


Figure 3: An illustrative scenario under hidden confounding, where the CAPO intervals for two actions overlap significantly.

An Illustrative Comparison

To highlight the advantage of our probabilistic causal target, we compare our CTLD framework to cost-sensitive methods, such as CARED (Ghoummaid and Shalit 2024), in a hidden confounding scenario. Assume for a patient with covariates x, the estimated CAPO intervals are $I_0 = [5, 15]$ and $I_1 = [8, 18]$, as depicted in Figure 3. Suppose an expert chose action $a = 1$, but we observe two different outcomes due to unobserved confounders: a high outcome $y_1 = 17$ and a modest one $y_2 = 9$.

Cost-Sensitive Approach (e.g., CARED). This approach's deferral cost, $C_\perp$, directly depends on the observed outcome y. For our two realizations, this cost changes drastically from $C_\perp(y_1) = 5 - 17 = -12$ to $C_\perp(y_2) = 5 - 9 = -4$. This demonstrates a key weakness: the learning signal is highly sensitive to outcome uncertainty, which can impair the learned policy's robustness.

Our CTLD Approach. In contrast, our method constructs a causal target directly from the interval geometry, making it independent of the noisy outcome.

- The **deferral probability**, determined by the intervals' *relative overlap*, is $\tilde{p}_{\text{defer}}(x) = |I_0 \cap I_1|/|I_0 \cup I_1| \approx 0.54$.
- The **action probabilities**, derived from *uncertainty-scaled* CATE scores, are calibrated to $\tilde{p}_{\text{action}}(x) \approx [0.39, 0.61]$.

The resulting causal target, $\tilde{p}(x) \approx [0.39, 0.61; 0.54]$, is **identical** for both y_1 and y_2. This principled decoupling from outcome uncertainty provides a stable learning target, which is a central advantage of our CTLD framework. Detailed calculations are provided in Appendix D.

To summarize our CTLD, Algorithm 1 presents the procedure in three high-level steps: estimating sharp bounds, constructing causal targets, and learning the L2D policy.

Algorithm 1: Causal-target-based learning to defer (CTLD)

Input: Dataset $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^N$, confounding level Λ, loss balance λ, model class $\mathcal{G}$
Output: Learned policy $\hat{\pi}(x)$
// Step 1: Estimate sharp bounds
Estimate sharp CAPO bounds $\hat{Q}^{\pm}(X_i, a)$ for all samples $i = 1, \ldots, N$.
// Step 2: Construct causal targets
for each X_i **do**
    Compute deferral probability $\tilde{p}_{\text{defer}}(X_i)$
    Compute action probabilities $\tilde{p}_{\text{action}}(X_i)$
    Form the complete target $\tilde{p}(X_i)$
end for
// Step 3: Learn L2D policy
Learn $g^* \in \mathcal{G}$ by minimizing total loss over all $\tilde{p}(X_i)$
return $\hat{\pi}(x)$ by Equation (10)

Theoretical Guarantees

Our theoretical analysis is presented in three interconnected parts. First, we establish the **consistency of our CTLD's core components** (Propositions 1 and 2), which confirms that our novel causal target is not an arbitrary construct but a statistically reliable learning target that converges to a meaningful quantity. Second, we present our main theoretical contribution: a **regret-transfer bound** (Theorem 1). This is the crucial bridge that formally connects minimizing our surrogate learning objective to the ultimate goal of minimizing the true, unobservable causal regret. Finally, we provide a standard **generalization bound** (Theorem 2) to ensure that the policy learned on the training data will perform reliably on unseen data. Collectively, these results provide a comprehensive theoretical validation, showing that CTLD is not only learnable and stable, but also a principled and effective approach to decision-making under uncertainty.

Assumption 2 (Regularity, Flexibility, and Convergence)**.** *We assume the following: (a)* ***Bounded Outcomes****:* $|Y| \leq C_Y$*. (b)* ***Sieve Approximation****: optimal scoring function* g^* *can be increasingly well-approximated by a sequence of function classes* $\{\mathcal{G}_m\}_{m\geq 1}$*, where* m *denotes the model complexity. Formally,* $\inf_{g\in\mathcal{G}_m} \|g - g^*\|_{L_2} = \mathcal{O}(m^{-\beta})$ *for some* $\beta > 0$*. (c)* ***Complexity Control****: For every* m*,* $\mathcal{R}_n(\mathcal{G}_m) = \mathcal{O}\left(\sqrt{\frac{\log m}{n}}\right)$*. (d)* ***Data-Driven Selection****: Choose* $m = m(n)$ *so that* $\sqrt{\frac{\log m(n)}{n}} \asymp n^{-\alpha}$ *with* $\alpha < \beta/(1+2\beta)$*. (e)* ***Bound Estimation Rate****:* $\|\widehat{Q}_n^{\pm} - Q^{\pm}\|_{L_2} = \mathcal{O}_p(n^{-\alpha})$*.*

Assumption (a) guarantees bounded potential outcomes and hence bounded sharp bounds. For (b) and (c), we employ a regularised sieve of weight–decayed ReLU networks $G_{m(n)}$ whose width grows as $m(n) \propto n^{1/(1+2\beta)}$. This construction (i) approximates any β–Hölder target at rate $m(n)^{-\beta}$, and (ii) enjoys a Rademacher complexity $\mathcal{R}_n(\mathcal{G}_{m(n)}) = \mathcal{O}\big(\sqrt{\log m(n)/n}\big)$ due to the spectral-norm constraint (Schmidt-Hieber 2020). We determine $m(n)$ via cross–validated early stopping, a practical method intended to approximate the balanced rate required by (d). Assumption (e) is justified by the established $n^{-\alpha}$ rate for the deep neural network nuisance estimators used in their construction (Farrell, Liang, and Misra 2021, Theorem 1). This architecture and training protocol are consistent with state–of–the–art implementations (Ghoummaid and Shalit 2024; Hess et al. 2025).

Proposition 1 (Consistency of Bound Estimators)**.** *Assumption 2(e) implies that the CAPO bound estimators* $\hat{Q}_n^{\pm}(x, a)$ *are consistent for the true sharp bounds* $Q^{\pm}(x, a)$*.*

See Appendix C for the full proof. This proposition ensures that the inputs to our causal target construction are asymp-

totically correct.

Proposition 2 (Consistency of the Causal Target)**.** *Under the consistency of the causal bound estimators, as the sample size $n \to \infty$, the causal target $\tilde{p}_n(X)$ converges in probability to a deterministic limiting vector $\tilde{p}^*(X)$ that correctly reflects the underlying causal uncertainty.*

See Appendix C for the full proof.

With these consistency propositions, our main theoretical contribution is a **regret-transfer bound**. Our ultimate goal is to find a policy that minimizes the true **causal regret**, which measures the performance gap to the unknown optimal policy. However, this objective cannot be optimized directly. Instead, during training, we minimize a tractable **true surrogate risk**, denoted $R(g)$, which is the expected cross-entropy loss against our causal targets. Let R^* be the minimum possible value of this risk. The quantity $R(g) - R^*$ is then the *excess surrogate risk*—a measure of how suboptimal our learned model is on the training objective. The following theorem provides the crucial bridge connecting causal regret to the surrogate risk.

Theorem 1 (Regret-Transfer Bound)**.** *For every score function g, let π_g be the policy derived from it and π^* be the oracle policy defined by the causal target. Then the excess causal regret is bounded by the excess surrogate risk:*

$$\mathbb{E}_X\big[\mathrm{Reg}_X(\pi_g)\big] - \mathbb{E}_X\big[\mathrm{Reg}_X(\pi^*)\big] \;\le\; \frac{U_{\max}}{\sqrt{2}}\sqrt{R(g) - R^*}.$$

The full proof is provided in Appendix C. The significance of this theorem is that it justifies our entire learning procedure. It provides a formal guarantee that by finding a model with a low training error (i.e., minimizing the surrogate risk $R(g)$), we are indeed making progress towards the true goal of finding a policy with low real-world regret.

Finally, we provide a finite-sample **generalization bound** for our CTLD. This result is crucial as it connects the empirical risk $\hat{R}(g)$ (what we minimize on our training data) to the true risk $R(g)$ (the performance on the real-world data distribution), guaranteeing that the policy learned on the training data will perform reliably on unseen data.

Theorem 2 (Generalization Bound)**.** *Recall that $R(g)$ is the **true surrogate risk** and let $\hat{R}(g)$ be the empirical surrogate risk. Under Assumption 2, for the empirical risk minimizer $\hat{g}_n$, with probability at least $1 - \delta$, the following holds:*

$$R(\hat{g}_n) - \hat{R}(\hat{g}_n) \le C_1 \mathcal{R}_n(\mathcal{G}) + C_2 n^{-\alpha} + C_3 \sqrt{\frac{\log(1/\delta)}{n}}, \tag{11}$$

where $\mathcal{R}_n(\mathcal{G})$ is the Rademacher complexity of the function class $\mathcal{G}$, and C_1, C_2, C_3 are constants.

The full proof is provided in Appendix C. This bound provides insight into the sources of generalization error, which consists of three key components: a term for the complexity of the model class ($\mathcal{R}_n(\mathcal{G})$), an approximation error term from using an estimated target ($n^{-\alpha}$), and a standard statistical learning term.

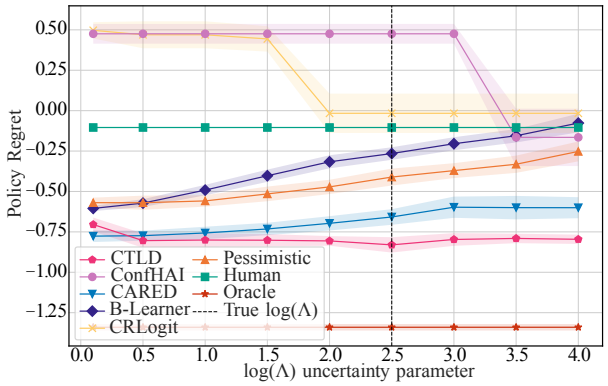


Figure 4: Policy regret on synthetic data. The x-axis shows the assumed level of hidden confounding ($\log(\Lambda)$), and lower policy regret indicates better performance. The vertical dashed line marks the true confounding level ($\log(\Lambda_0) = 2.5$). Our method, CTLD, consistently outperforms key baselines across nearly all levels of confounding.

Experiments

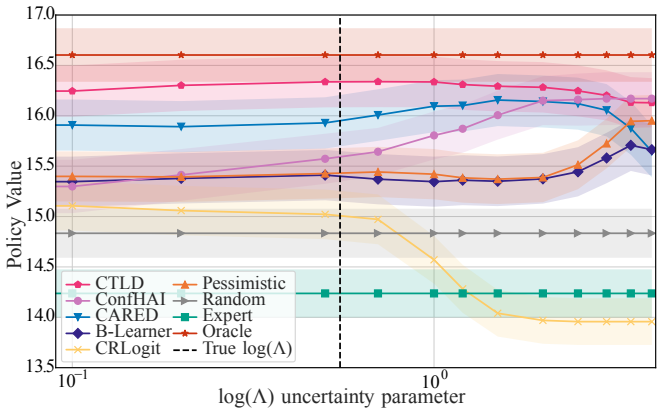
We conduct experiments to evaluate the effectiveness of our proposed method, **CTLD**, in learning deferral policies under hidden confounding. We use two experimental settings: (1) a synthetic environment adapted from (Gao and Yin 2025) to allow for controlled analysis, and (2) a semi-synthetic setup based on the IHDP dataset (Hill 2011) to assess performance on more realistic data distributions. In both settings, we vary the marginal sensitivity parameter Λ to simulate different degrees of unobserved confounding.

Baselines and Setup

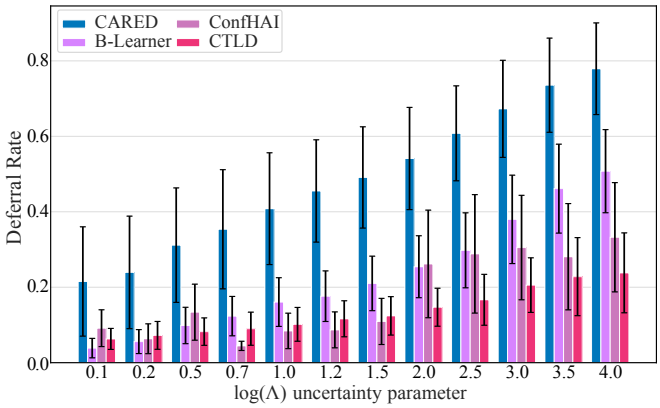
We compare our **CTLD** against a comprehensive range of recent and standard baselines:

- **ConfHAI** (Gao and Yin 2025): A minimax IPW approach for deferral policy learning.
- **CARED** (Ghoummaid and Shalit 2024): A cost-sensitive classification method for deferral policy learning.
- **B-Learner** (Oprescu et al. 2023; Jesson et al. 2021): A policy that defers when CATE interval bounds cross zero, otherwise assigns treatment.
- **CRLogit** (Kallus and Zhou 2021): A minimax policy learning model without any deferral mechanism.
- **Pessimistic Policy**: A conservative policy that avoids deferral by choosing the action with the better worst-case.
- **Random Deferral Policy**: Randomly defers with fixed probability, serving as a naive baseline.
- **Human/Expert Policy**: A human-derived policy, simulating human decision-makers recommendations.
- **Oracle**: A clairvoyant policy with access to true potential outcomes, used as an upper bound.

Comprehensive details regarding the experimental setup, data generation, and hyperparameter settings for all models, are provided in Appendix E.



(a) Policy value for different values of $\log(\Lambda)$.



(b) Deferral rate for different values of $\log(\Lambda)$.

Figure 5: Performance on the semi-synthetic IHDP dataset. (a) Policy value (higher is better) versus the assumed confounding level $\log(\Lambda)$, with the true value marked by a vertical dashed line. (b) Corresponding deferral rates for the deferral-enabled policies. CTLD achieves the highest policy value among learned policies while employing a highly efficient deferral strategy, with a deferral rate consistently less than a third of that of the CARED baseline.

Synthetic Dataset

We replicate the synthetic data experiment from (Gao and Yin 2025) and run 10 trials. For each trial, we vary the uncertain parameter $\log(\Lambda)$ from 0.1 to 4, corresponding to various levels of assumed hidden confounding. We compare the *policy regret* for the returned policy for each method relative to the Baseline Policy, which assigns action $a = 0$ for all individuals. Additionally, we report the policy regret of Human Policy. Figure 4 shows the results. CTLD achieves the lowest regret among nearly all learned policies, closely approaching the Oracle performance. Notably, the performance of methods like ConfHAI and CARED deteriorates significantly when the assumed confounding level $log(\Lambda)$ is misspecified (i.e., far from the true value of 2.5). In contrast, CTLD remains stable across the entire range, demonstrating robustness to the misspecification of the confounding level.

IHDP Dataset

To evaluate performance in a more realistic setting, we test our method on the semi-synthetic IHDP dataset, following the setup from (Jesson et al. 2021). The results, averaged over 200 dataset realizations, are presented in Figure 5.

The findings highlight the effectiveness and efficiency of our CTLD framework. As shown in Figure 5(a), CTLD consistently achieves the highest policy value among nearly all learned policies, robustly outperforming other methods across the entire range of confounding levels. Simultaneously, Figure 5(b) demonstrates that this superior performance is achieved with remarkable efficiency, as CTLD's deferral rate is consistently less than a third of that required by the CARED. This result empirically validates that our core proposal of constructing the probabilistic learning target leads to a more effective and efficient deferral policy.

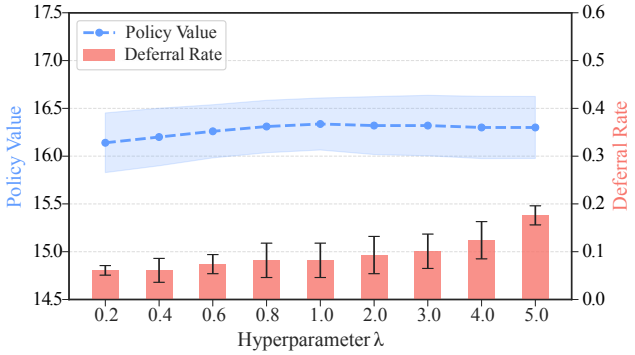


Figure 6: Ablation study on the hyperparameter λ in Equation (7). Larger λ increases the deferral rate while keeping the policy value stable.

Ablation Study

To assess the effect of λ, we conduct an ablation at $\log(\Lambda) = 0.5$ and report results averaged over 200 trials. As shown in Figure 6, increasing λ monotonically raises the deferral rate. Crucially, this is achieved while the policy value remains high, indicating that CTLD learns to defer on genuinely uncertain cases without sacrificing performance.

Conclusion

In this paper, we study learning to defer under hidden confounding using only observational data. We introduce CTLD framework, which constructs a stable probabilistic target from sharp bounds on potential outcomes. CTLD is supported by regret-transfer and generalization guarantees. Experiments show that CTLD achieves strong policy value with an efficient deferral rate, remaining robust even under misspecified confounding.

Acknowledgments

We thank all the anonymous reviewers and meta reviewers for their valuable comments, as well as all of our team members for their support and assistance.

References

Athey, S.; and Wager, S. 2021. Policy learning with observational data. *Econometrica*, 89(1): 133–161.

Bondi, E.; Koster, R.; Sheahan, H.; Chadwick, M.; Bachrach, Y.; Cemgil, T.; Paquet, U.; and Dvijotham, K. 2022. Role of human-AI interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5286–5294.

Cao, Y.; Mozannar, H.; Feng, L.; Wei, H.; and An, B. 2023. In defense of softmax parametrization for calibrated and consistent learning to defer. *Advances in Neural Information Processing Systems*, 36: 38485–38503.

Dorn, J.; and Guo, K. 2023. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544): 2645–2657.

Farrell, M. H.; Liang, T.; and Misra, S. 2021. Deep neural networks for estimation and inference. *Econometrica*, 89(1): 181–213.

Gao, R.; Saar-Tsechansky, M.; De-Arteaga, M.; Han, L.; Lee, M. K.; and Lease, M. 2021. Human-AI Collaboration with Bandit Feedback. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 1722–1728.

Gao, R.; and Yin, M. 2025. Confounding-robust deferral policy learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14238–14246.

Ghoummaid, M.; and Shalit, U. 2024. When to act and when to ask: policy learning with deferral under hidden confounding. *Advances in Neural Information Processing Systems*, 37: 56108–56135.

Hemmer, P.; Thede, L.; Vössing, M.; Jakubik, J.; and Kühl, N. 2023. Learning to defer with limited expert predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6002–6011.

Hess, K.; Frauen, D.; Melnychuk, V.; and Feuerriegel, S. 2025. Efficient and sharp off-Policy learning under unobserved confounding. arXiv:2502.13022.

Hill, J. L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1): 217–240.

Ho, K.; and Rosen, A. M. 2015. Partial identification in applied research: benefits and challenges. Technical report, National Bureau of Economic Research.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Jesson, A.; Mindermann, S.; Gal, Y.; and Shalit, U. 2021. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, 4829–4838. PMLR.

Kallus, N. 2018. Balanced Policy Evaluation and Learning. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kallus, N.; and Zhou, A. 2018. Confounding-robust policy improvement. *Advances in neural information processing systems*, 31.

Kallus, N.; and Zhou, A. 2021. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5): 2870–2890.

Leitão, D.; Saleiro, P.; Figueiredo, M. A. T.; and Bizarro, P. 2022. Human-AI Collaboration in Decision-Making: Beyond Learning to Defer. arXiv:2206.13202.

Mozannar, H.; Lang, H.; Wei, D.; Sattigeri, P.; Das, S.; and Sontag, D. 2023. Who should predict? exact algorithms for learning to defer to humans. In *International conference on artificial intelligence and statistics*, 10520–10545. PMLR.

Mozannar, H.; and Sontag, D. 2020. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, 7076–7087. PMLR.

Oprescu, M.; Dorn, J.; Ghoummaid, M.; Jesson, A.; Kallus, N.; and Shalit, U. 2023. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *International Conference on Machine Learning*, 26599–26618. PMLR.

Ruggieri, S.; and Pugnana, A. 2025. Things machine learning models know that they don't know. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28684–28693.

Schmidt-Hieber, J. 2020. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4): 1875–1897.

Tailor, D.; Patra, A.; Verma, R.; Manggala, P.; and Nalisnick, E. 2024. Learning to defer to a population: A meta-learning approach. In *International Conference on Artificial Intelligence and Statistics*, 3475–3483. PMLR.

Tan, Z. 2006. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476): 1619–1637.