Analyzing the Impact of Depth and Leaf Size on CATE Estimation in Honest Causal Trees

A Study of Model Accuracy and Generalization Across Simulated and Real-World Data

Rheea-Maria Prodan
Supervisor(s): Jesse Krijthe, Rickard Karlsson
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Rheea-Maria Prodan
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijther, Rickard Karlsson, Ricardo Marroquim

Abstract

Causal inference, particularly the estimation of the Conditional Average Treatment Effects (CATE), is necessary for understanding the impact of interventions beyond simple predictions. This study analyzes the influence of key hyperparameter choices, specifically maximum tree depth and minimum leaf size, on the accuracy and generalization of CATE estimates derived from honest and adaptive causal trees. The research explores how these hyperparameters affect the bias-variance trade-off and the model's tendency to overfit or underfit across various simulated data scenarios and a real-world dataset.

The results reveal that optimal hyperparameter configurations are dependent on the data characteristics, such as dimensionality, noise levels, and the complexity of the true causal effects. Honest causal trees demonstrate a better performance in high-dimensional and noisy environments due to their effective variance control. Conversely, in simpler, low-noise settings or complex CATE structures, adaptive causal trees or baseline models frequently achieve better results by reducing bias. The study also highlights the challenges of using moderately sized datasets, where the sample splitting limitations can lead to higher estimation errors. This work provides thorough suggestions for hyperparameter selection, emphasizing the fact that tuning based on the underlying characteristics of the data is needed for achieving the best CATE estimates possible.

1 Introduction

In standard machine learning algorithms, the focus is on predicting an outcome based on a set of observed inputs, typically by learning statistical associations from data. For example, one might use the patient's age or blood test results to predict their likelihood of developing a certain disease. In contrast, causal machine learning aims to go further than forecasting by understanding the direct effect of interventions, in other words, what would happen to an outcome if a variable were changed, such as administering a treatment [1]. For instance, instead of merely associating treatment with recovery, causal inference seeks to estimate the effect of receiving a treatment versus not receiving it. Moreover, this effect may vary across individuals. A central quantity in this context is the Conditional Average Treatment Effect (CATE), which captures the expected effect of a treatment conditional on the individual's characteristics, enabling personalized decision-making.

Still, causal inference poses unique challenges. Specifically, for each individual, only one of the potential outcomes, what they would experience if treated versus if untreated, is observed, making the counterfactual unobservable. To estimate causal effects from observational data, key assumptions such as the Stable Unit Treatment Value Assumption (SUTVA), positivity (or overlap), and unconfoundedness are required. These assumptions are crucial for identifying causal effects from data [2].

In order to address the challenge of estimating causal effects rather than mere associations, adaptations of already known and used machine learning models are required. A prominent approach in this domain is the honest estimation framework, as proposed by Athey and Imbens [3], specifically designed for causal decision trees. In contrast to adaptive decision trees, which reuse the same data for both tree construction and effect estimation, honest trees split the data into separate subsamples for splitting and estimation. This approach helps reduce overfitting by mitigating selection bias that could occur when reusing the same data for both tree construction and effect estimation, thus ensuring that splitting decisions do not mistakenly incorporate or amplify random fluctuations from the data into the treatment effect estimates.

However, as in any other machine learning model, the performance of causal decision trees is influenced by hyperparameter choices. In particular, maximum tree depth and minimum leaf size can significantly affect the model's bias-variance trade-off, influencing both the accuracy of CATE estimates and the model's tendency to overfit or underfit. This raises the central research question: how do key hyperparameter choices, specifically maximum tree depth and minimum leaf size, affect the accuracy and tendency to overfit or underfit in CATE estimates produced by honest causal decision trees, across both simulated and real-world data settings?

To address this, the study explores these conditions by isolating the effects of each hyperparameter, namely by systematically varying the maximum tree depth and the minimum leaf size across diverse simulated and real-world data environments. The primary contribution of this research is to establish practical guidelines for hyperparameter selection in honest causal trees, concluding that optimal configurations are dependent on the underlying characteristics of the data, such as dimensionality, noise levels, the complexity of the CATE function, as well as the size of the dataset. This work offers suggestions for hyperparameter selection, revealing that while both honest and adaptive implementations demonstrate robustness against severe overfitting, a nuanced bias-variance trade-off is observed based on the data context. Thus, the optimal configurations often involve shallower trees with larger leaves in noisy or high-dimensional settings, while deeper trees with smaller leaves are better suited for complex or low-noise CATEs.

The remainder of this paper continues with section 2, which lays the background for the causality fundamentals, identification assumptions, and the honest framework, as well as introducing the existing gap. In section 3, the methodology, including data sources, experiment setup, and evaluation metrics, is detailed. Next, section 4 presents the results from both simulated and real-world data analyses. In section 5, a discussion of the findings,

which outlines the limitations and suggests future research directions, is offered. Following this, section 6 reflects on the ethical implications, reproducibility and transparency of this research. Lastly, section 7 concludes the paper with a summary of key insights.

2 Background

This section dives deeper into the existing research and theories that form the foundation of the research question. In subsection 2.1, the potential outcome framework and Conditional Average Treatment Effect (CATE) are introduced. Following this, subsection 2.2 lays out the three core identification assumptions, namely SUTVA, overlap, and unconfoundedness. In subsection 2.3, the honest causal tree framework of Athey and Imbens is defined, highlighting how sample splitting and modified split criteria yield unbiased CATE estimates with valid inference. Finally, subsection 2.4 discusses the lack of systematic guidance on tuning key hyperparameters in honest causal trees, and how this gap motivates the research conducted.

2.1 Causality Fundamentals

Causal effects are defined via the potential-outcome framework, which states that each unit i has two potential outcomes: $Y_i(1)$ if treated and $Y_i(0)$ if untreated [2]. The individual treatment effect (ITE) for unit i is then $Y_i(1) - Y_i(0)$. However, the fundamental problem of causal inference arises from the existence of counterfactuals, or outcomes under the unobserved treatment. Specifically, the issue is that for any given unit, only one potential outcome can be observed, while the other remains unobserved. In empirical studies, an outcome variable Y_i, a binary treatment assignment $W_i \in \{0, 1\}$ which indicates whether the unit received treatment or control, and a vector of individual characteristics of covariates X_i are observed. Consequently, causal inference often focuses on estimands such as the Conditional Average Treatment Effect (CATE), $\tau(x) = E[Y(1) - Y(0)|X = x]$, which quanitifies how the treatment effect varies across different covariate profiles X, thus enabling the estimation of heterogeneous and personalized treatment effects.

2.2 Identification Assumptions

Estimating causal effects from observational data, where treatment assignment is not controlled or randomized, requires a set of assumptions that connect the observed data to the counterfactual estimands. One such assumption is the Stable Unit Treatment Value Assumption (SUTVA), which has two components, namely no interference between units, so that the potential outcome for each unit depends only on its own treatment, and well-defined treatments so that there is a single version of each treatment level [4]. Another assumption is that of overlap, or positivity, which requires that every unit has a nonzero probability of receiving each treatment level. If the overlap assumption fails, estimates become extrapolations and can be highly unstable [5]. Finally, one vital assumption is that of unconfoundedness or ignorability, which states that all confounders are observed and captured in the feature set [6]. A confounder is a variable that influences both the treatment assignment and the outcome, creating an association that can obscure the true causal effect if not accounted for. Under these assumptions, the CATEs are identifiable from observed data via methods such as propensity-score weighting, matching, or outcome modelling, while violations can introduce bias or invalidate inferential guarantees, meaning that statistical claims about the precision and reliability of estimates may no longer be accurate [6].

2.3 Honest Framework

Athey and Imbens introduce the honest causal tree, a modification of decision trees, specifically designed to address challenges inherent in causal inference [3]. Unlike standard regression trees, which predict outcomes by recursively partitioning data to minimize outcome variance or mean squared error with terminal nodes and reuse the same data both to select splits and to estimate leaf predictions, thereby inducing selection bias, the honest causal tree employs a sample splitting strategy. Selection bias arises when the same data used to discover a split is also used to estimate the effect in the resulting groups. This can lead to splits based on random noise in that specific sample rather than true heterogeneity.

In contrast, the adaptive variant of causal trees does not employ such a sample splitting strategy. It reuses the same data for both determining the tree structure and estimating the treatment effects within the leaf nodes. This approach allows adaptive trees to utilize the full dataset size for both stages, which can come at the risk of increased overfitting and bias, as the model can learn spurious correlations from the training data when making the splitting decisions and estimating effects from the same samples [7].

To mitigate this, the honest causal tree partitions the available data into two disjoint subsets, one of which is used to determine the tree structure, identifying covariates and thresholds for splits, and the other one is used exclusively to compute unbiased treatment effect estimates within each leaf node. This separation

mitigates selection bias that would otherwise arise from data reuse, preventing random noise from influencing the effect estimates. The splitting criterion itself is further modified to explicitly optimize for treatment-effect heterogeneity rather than simply minimizing outcome purity.

Recognizing that the true unit-level causal effect is never observed, Athey and Imbens propose unbiased plug-in estimators for both the empirical Mean Squared Error (MSE) of the estimated treatment effects on the splitting sample and the sample variance of the leaf-specific treatment effect. These estimators are derived under the potential outcomes framework [2] and are crucial because their formulation accounts for the fact that each leaf's effect will be estimated from an independent hold-out sample. This design choice makes the splitting decisions identify true heterogeneity, rather than overfitting to noise in the splitting sample.

To select the overall tree complexity, the authors develop a tailored cross-validation procedure. Each fold again splits into train-split and train-estimation subsamples, and performance is evaluated in terms of out-of-sample CATE MSE rather than traditional outcome MSE. Simulation studies in the paper demonstrate that honest trees achieve nominal coverage for 90% confidence intervals of CATEs, whereas non-honest trees under-cover substantially.

Finally, the honest splitting strategy manages the bias-variance trade-off, thus honesty effectively reduces selection bias by separating data for splitting and estimation. However, this comes at the cost of higher variance due to smaller effective sample sizes available for estimation within each leaf. Despite this increased variance, the method typically yields a lower overall MSE for treatment effect estimates in moderate to large samples.

2.4 Research Gap and Motivation

While hyperparameter tuning is well recognized as critical for controlling bias-variance trade-offs in standard tree algorithms, its role in causal tree models, especially honest causal trees, remains underexplored. In supervised learning, maximum tree depth and minimum leaf size directly influence model complexity: deeper trees with small leaves tend to overfit, whereas shallow trees with large leaves may underfit, being too simplistic to capture underlying patterns, resulting in high bias [8].

By contrast, causal effect estimation imposes additional considerations. Causal-attributed splits must balance capturing effect heterogeneity against incurring excessive estimation variance in hold-out samples. The literature on "Hyperparameter tuning and model evaluation in causal effect estimation" highlights that modern causal estimators rely on highly tuned hyperparameters to avoid large estimation errors in high-dimensional settings [9]. Yet, their analysis focuses on general causal effect estimation methods rather than specifically on the honest causal tree framework introduced by Athey and Imbens [3], so there remains no empirical roadmap for selecting tree-specific hyperparameters. This absence leaves practitioners without clear recommendations for selecting hyperparameters that balance heterogeneity detection and honest estimation variance control.

To fill this gap, this research conducts a comprehensive study that investigates how maximum tree depth and minimum leaf size affect CATE estimation accuracy, overfitting, and underfitting in honest causal trees, using both simulated benchmarks and real-world datasets.

3 Methodology

This section details the research design and the specific procedures used to conduct the study. In subsection 3.1 the synthetic data generation processes, including the established Athey and Imbens designs and custom-tailored DGPs, and the use of real-world data are explained. Next, subsection 3.2 describes the experiment configurations, specifying the models employed, the hyperparameters varied, and the replication strategies. This is followed by subsection 3.3, which elaborates the quantitative metrics used for evaluating estimator performance. Finally, subsection 3.4 introduces the expected results from the experiments.

3.1 Data Sources

To evaluate the impact of hyperparameter choices on the performance of honest causal trees, both simulated datasets and a real-world benchmark dataset are used. This combination allows for controlled experimentation alongside practical validation.

3.1.1 Simulated Data

Athey and Imbens have proposed three data simulation designs to create controlled environments for assessing estimator performance [3]. These simulations have the purpose of presenting varying degrees of treatment effect heterogeneity and complexity, providing insights into how honest causal trees respond to different data-generating processes. All designs generate an outcome Y_i, a K-component vector of features X_i, a binary treatment indicator $W_i \in \{0, 1\}$, and the true causal effect τ_i. The observed outcome is given by $Y_i = \eta(X_i) + \frac{1}{2}(2W_i - 1)\tau(X_i) + \epsilon_i$, where $\eta(X_i)$ is the mean outcome function, $\tau(X_i)$ is the conditional average treatment

effect (CATE) function, and ϵ_i is a noise term, with $X_i \sim N(0,1)$ and $W_i \sim Bernoulli(0.5)$. For all the designs $\epsilon_i \sim N(0, 0.1)$.

It is important to note that for all simulated designs, the binary treatment indicator W_i is generated as $W_i \sim Bernoulli(0.5)$, which implies a randomized treatment assignment. Therefore, these simulated datasets inherently assume no observed or unobserved confounding variables influencing both treatment assignment and outcome.

Design 1: Simple, Low Dimensionality
The first design, and the simplest one, has two covariates. $X_{i,0}$ affects both the mean outcome and the treatment effect, while $X_{i,1}$ affects only the mean outcome function. This means that there are no irrelevant covariates, allowing for a straightforward assessment of the estimator's performance in an ideal setting.
Design 2: Moderate Dimensionality with Non-informative of Treatment Effect Covariates
The second design adds more complexity by having 10 covariates (K=10), out of which only two of them affect the treatment effect (X_0, X_1), conditional on their values being positive. The remaining eight covariates are non-informative of the treatment effect, introducing moderate complexity and testing the estimator's ability to differentiate relevant variables.
Design 3: High Dimensionality with Non-informative of Treatment Effect Covariates
Lastly, the third data generation design, and the most complex one, presents 20 covariates (K=20), out of which only four, if positive, affect the treatment effect, the rest of them representing noise covariates. This creates a high-dimensional environment and aims to challenge the estimator's capacity to identify relevant variables.

3.1.2 Custom Data Generation Processes

To further explore the impact of specific data characteristics beyond the Athey & Imbens designs, four additional data generation processes (DGPs) are introduced. These designs isolate factors such as varying dimensionality of relevant covariates, non-linear treatment effects, interaction terms, and different levels of noise. For all custom designs $\epsilon_i \sim N(0, \sigma^2)$, where σ is specified for each design.
Design 4: Varying Dimensionality, All Relevant Covariates
The fourth data generation design is motivated by the observation that Athey and Imbens' first design might be too simple to induce prominent overfitting behaviour in adaptive trees. To test the impact of dimensionality when all covariates directly affect the treatment effect, the fourth DGP ensures all features contribute to $\tau(X_i)$ linearly, while $\eta(X_i)$ remains simple. This allows for a clear assessment of how estimators handle high-dimensional heterogeneity without confounding factors of irrelevant variables. The sensible values for K chosen are 2, 5, 10, 15, 20, 25, 50, with $\sigma = 0.1$.
Design 5: Non-Linear CATE
The fifth DGP is designed to isolate the impact of non-linearity in the treatment effect function. Unlike Athey and Imbens' designs, which use piecewise linear non-linearity, this introduces a continuous, oscillating non-linear relationship. The use of sinusoidal functions is preferred over polynomials here because they inherently introduce non-monotonicity and periodicity, which are observable within the typical range X takes from a standard normal distribution. For example, Hyvärinen, Shimizu, and Hoyer demonstrate that incorporating sinusoidal components can reveal non-monotonic causal relationships that are difficult to capture with standard polynomial models [10]. The mean outcome function remains simple, and no noise covariates are present, allowing for a focused analysis of how trees capture inherent non-linear heterogeneity. Sensible values for this are 2, 5, while $\sigma = 0.1$.
Design 6: CATE with Interaction Terms
This design aims to study the estimator's ability to uncover interaction effects within the treatment effect, which are needed for identifying specific subgroups where treatment effects might be amplified or diminished. Multiplicative interaction terms create a conditional change in effect, making it different from simple adaptive heterogeneity. Studying interaction terms is valuable because they reveal subgroups where the treatment effect differs, enabling more precise and personalized causal conclusions [11]. $\eta(X_i)$ is kept simple, $\sigma = 0.1$, and there are no noise covariates present, to ensure that the focus remains on differentiating the multiplicative relationship between covariates.
Design 7: Varying Noise Levels
This DGP aims to systematically evaluate the robustness of honest and adaptive causal trees under different levels of noise. By controlling the standard deviation term σ, this design assesses how increasing uncertainty in the outcome measurements affects the accuracy and stability of CATE estimates, and whether honesty provides particular advantages in noisy environments.

The characteristics and formulas for all data generation processes are summarized in Table 1.

Table 1: Summary of Data Generating Processes (DGPs)

Design ID	Description	K (Features)	Noise σ (Type)	$\eta(X_i)$ Formula	$\tau(X_i)$ Formula
Athey and Imbens Designs					
1	Simple, Low Dim	2	0.1	$0.5 \cdot X_{i,0} + X_{i,1}$	$0.5 \cdot X_{i,0}$
2	Moderate Dim, Non-informative of Treatment Effect Covariates	10	0.1	$0.5 \sum_{k=0}^{1} X_{i,k} + \sum_{k=2}^{5} X_{i,k}$	$\sum_{k=0}^{1} \mathbf{1}\{X_{i,k} > 0\} \cdot X_{i,k}$
3	High Dim, Non-informative of Treatment Effect Covariates	20	0.1	$0.5 \sum_{k=0}^{3} X_{i,k} + \sum_{k=4}^{7} X_{i,k}$	$\sum_{k=0}^{3} \mathbf{1}\{X_{i,k} > 0\} \cdot X_{i,k}$
Custom Designs					
4	Varying Dim, All Relevant Covariates	2, 5, 10, 15, 20, 25, 50	0.1	$0.5 \cdot X_{i,0} + X_{i,1}$	$0.5 \sum_{k=0}^{K-1} X_{i,k}$
5	Non-Linear CATE	2, 5	0.1	$0.5 \cdot X_{i,0} + X_{i,1}$	$\sin(X_{i,0}) + 2 \cdot \cos(X_{i,1})$
6	CATE with Interaction Terms	2, 5	0.1	$0.5 \cdot X_{i,0} + X_{i,1}$	$X_{i,0} \cdot X_{i,1}$
7	Varying Noise Levels	2	$0.01, 0.1, 0.25, 0.5, 0.75, 1$	$0.5 \cdot X_{i,0} + X_{i,1}$	$0.5 \cdot X_{i,0}$

3.1.3 Real-World Data

For real-world validation, the Infant Health and Development Program (IHDP) dataset is used, a well-known benchmark in causal inference research [12]. The IHDP is a randomized controlled trial aimed at evaluating the effect of early educational and health interventions on the cognitive development of low birth weight, premature infants. This dataset has 747 observations with 25 covariates, making it a moderately sized benchmark.

A semi-synthetic version of the IHDP dataset is used, where the original treatment assignments are retained, but outcomes are simulated to introduce a known ground truth for treatment effects. This approach allows for the assessment of estimator accuracy in a realistic setting while maintaining the ability to compute true error metrics.

3.2 Experiment Setup

To conduct the experiments, EconML's implementation of causal trees was used, namely `CausalForest`, with the parameters `n_estimators=1` for having one tree and `honest=True` for enforcing the honest framework, or `honest=False` for adaptive trees [13].

To isolate the effect of the two hyperparameters, the experiments can be split into two main categories, one for the `max_depth` hyperparameter and one for the `min_samples_leaf` hyperparameter. To isolate the effects of each of these, whilst one is varied with values ranging from 2 to 20, the other keeps its default value, namely `max_depth=None` and `min_samples_leaf=5`. Additionally, for the real-world data setting, to verify the overall results of isolating each hyperparameter and to study how they behave when combined, a grid search across all combinations of `max_depth` and `min_samples_leaf` was performed.

For each experimental setting, the performance of both honest and adaptive versions of the causal tree estimator is evaluated. This allows for a direct comparison of how honesty influences the bias-variance trade-off and overall estimation accuracy across varying data complexities and hyperparameter configurations. Additionally, a simple T-Learner model will serve as a baseline for comparison. The T-Learner operates by fitting two independent regression models, one for the treated group and one for the control group. In this study, `DecisionTreeRegressor` models, with hyperparameters matched to the causal tree being evaluated, are used as the base learners for the T-Learner, to ensure a fair comparison and isolate the impact of the causal tree's specific splitting and estimation strategies. The CATE for any individual is then computed as the difference between the predictions from these two models. This provides a straightforward benchmark to assess whether the more complex tree-based approaches offer significant performance gains.

Simulated Data
When using synthetic data, each generated dataset had a sample size of 1000. To assess the performance of causal trees across different hyperparameter configurations, 100 Monte Carlo replications are employed [14]. This implies repeatedly simulating datasets under the same hyperparameter conditions and averaging the results to obtain stable estimates of metrics, such as bias, variance, and mean squared error of the estimated CATEs. The standard error of each averaged metric is also computed to provide a measure of its statistical precision. By running multiple simulations, the approximation of such metrics is more accurate, which is useful when assessing the performance of statistical estimators under different configurations. Thus, this ensures that the evaluation of the causal tree estimators was indeed robust, rather than influenced by random fluctuations in a single dataset.

Real-World Data
In real-world settings, the estimator's reliability is generalized via bootstrapped resampling [7]. Specifically, B=100 bootstrapped samples are drawn, sampled with replacement from the IHDP dataset, for which both the honest and adaptive versions of the causal trees are fitted on each sample, while keeping the hyperparameters

fixed. For each bootstrapped replicate, the out-of-sample CATE estimate is computed on the observations not selected in that replication, and the squared bias, variance, and mean squared error are recorded. Averaging these metrics over all B replications yields robust estimates of the sampling distribution of the CATE estimator under each hyperparameter configuration. The bootstrap method provides a non-parametric approach to approximating an estimator's sampling variability when analytical derivations are infeasible.

3.3 Evaluation Metrics

To assess the performance of the estimators within each experiment, the three core metrics of bias, variance, and mean squared error are used, all computed with respect to the true conditional treatment effect [15].

Bias is measured as the absolute difference between the average of these estimates across replications and the true effect. Its importance comes from its ability to quantify the systematic error by stating how far, on average, the estimated treatment effect strays from the true effects.

$$\text{Bias}(\hat{\tau}(x)) = E[\hat{\tau}(x)] - \tau(x)$$

Variance is the sample variance of the values over replications, which captures the stability of the estimation. This means that variance helps in understanding how much the estimates fluctuate from one sample to another.

$$\text{Var}(\hat{\tau}(x)) = E[(\hat{\tau}(x) - E[\hat{\tau}(x)])^2]$$

The Mean Squared Error (MSE) combines both of these components and provides a single measure. In this way, MSE penalizes both systematic and random errors, making it useful when tuning hyperparameters, as it highlights the classic bias-variance trade-off.

$$\text{MSE}(\hat{\tau}(x)) = E[(\hat{\tau}(x) - \tau(x))^2] = \text{Bias}(\hat{\tau}(x))^2 + \text{Var}(\hat{\tau}(x))$$

3.4 Expected results

For every experiment, the above-mentioned metrics will be plotted against all the values of hyperparameters considered. Every plot will also contain the standard error bands around the mean estimates, seen as a shaded area around each line. The expected result is a U-shaped curve, representing the bias-variance trade-off, meaning that initially increasing the hyperparameter will reduce bias by allowing the model to capture finer heterogeneity, but beyond a critical point, the variance will dominate, causing the MSE to increase [15]. However, it is hypothesized that honest causal trees will exhibit a flatter or less pronounced U-shaped MSE curve, compared to adaptive causal trees, as honesty is designed to control overfitting by mitigating bias from spurious correlations, particularly in more complex designs.

When varying the number of covariates without noise, it is anticipated that increasing dimensionality will generally lead to a higher MSE for both honest and adaptive causal trees, but honest trees should maintain better control over variance and bias, resulting in a more stable performance. For non-linear DGPs and the ones involving interaction terms, it is expected that the models will be more challenged and thus leading to a higher MSE. Lastly, increasing the noise level in the outcomes is expected to increase the MSE for both honest and adaptive causal trees.

4 Results

This section presents the findings from both simulation studies and real-world data validation, by systematically investigating how key hyperparameter choices, specifically maximum tree depth and minimum leaf size, influence the CATE estimation accuracy, overfitting, and underfitting in honest causal trees. These behaviours are compared against adaptive causal trees and a T-Learner baseline across various DGPs and the IHDP real-world dataset. In subsection 4.1, the findings from a series of controlled experiments are presented by investigating how hyperparameter choices impact the CATE estimation accuracy and generalization across diverse DGP configurations. Following this, subsection 4.2 assesses the model's performance in a real-world setting, by studying the effects of hyperparameters on the IHDP dataset, including an exploration of hyperparameter interactions.

4.1 Simulated Data

To establish a foundational understanding of honest and adaptive causal tree behaviours and to validate the simulation setup against established benchmarks, a series of preliminary experiments were conducted. These explorations provided crucial insights into the general performance and the bias-variance trade-off under various conditions, which have given the final selection of DGPs considered to be further analysis.

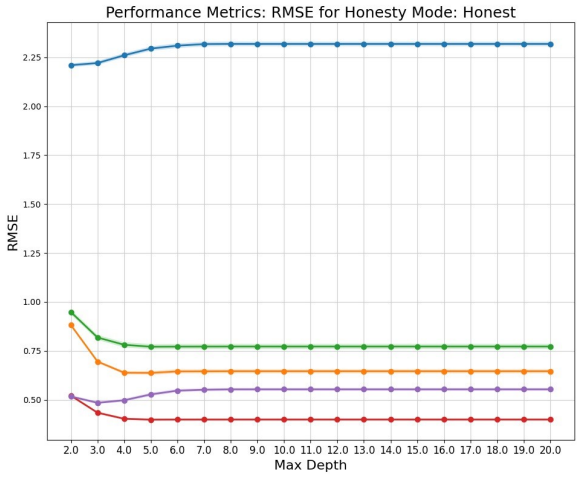
4.1.1 Comparison of Performance Across Selected DGPs

In order to synthesise the performance on honest causal trees, a subset of representative DGPs was considered. The selection of these DGPs aims to provide a holistic view of how hyperparameter choices of maximum depth and minimum leaf size influence the CATE estimation accuracy in honest causal trees, in comparison to adaptive causal trees and a T-Learner baseline. The first selected configuration was Design 4 with K = 20, as that was the value at which the honest causal trees started to outperform the adaptive causal trees when increasing the number of relevant covariates, results which can be found in Appendix B. In the context of adding additional complexity, by introducing non-linearity and interaction terms, Designs 5 and 6 have been chosen. The number of features has proved not to affect the MSE as described in Appendix C, thus an arbitrary value of K=5 has been chosen. Finally, to assess the performance of honest causal trees in different levels of noise, two values for the σ have been chosen. The first one is $\sigma = 0.1$, which corresponds to Design 1. The second one is $\sigma = 0.5$, a relatively large noise value, which interestingly has shown the expected U-shaped curve, as highlighted in Appendix D.

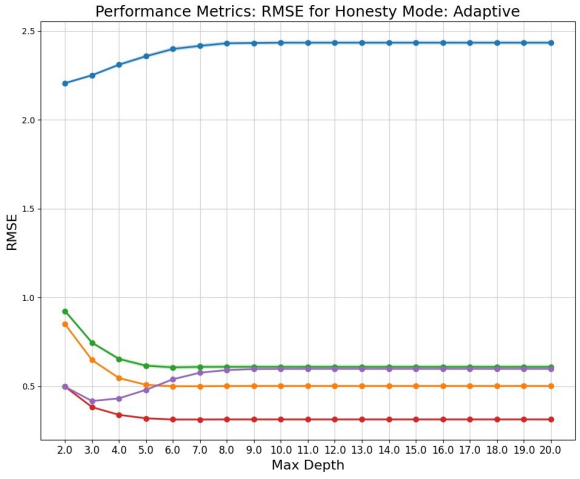
Maximum Depth

The impact of the maximum tree depth on MSE is highly dependent on the DGP's characteristics, as observed in Figure 1. For Design 5 (Non-Linear CATE), Design 6 (Interaction Terms), and Design 7 with $\sigma = 0.1$ (Low Noise Level), the MSE exhibits an initial decrease, after which it converges, indicating a robust control against severe overfitting where performance does not degrade at higher depths. This pattern suggests that for these DGPs, a certain tree depth is necessary to fully capture the underlying relationships.

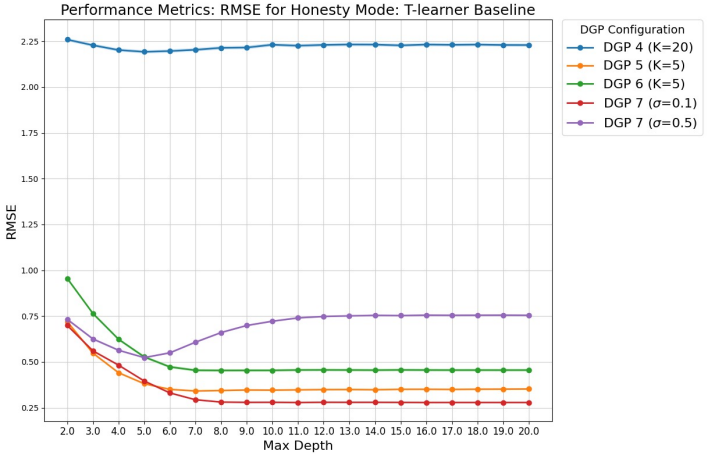
In contrast, for Design 4 (Medium-Dimensionality with Relevant Covariates) and for Design 7 with $\sigma = 0.5$ (Moderate Noise Level), the MSE curve shows an opposing behaviour. Here, the stabilization of the MSE occurs only after a continuous increase, suggesting that a smaller value of the `max_depth` hyperparameter helps in achieving a lower MSE. Moreover, for Design 7 with moderate noise, a more pronounced U-shaped curve is observed, confirming that higher noise levels can indeed induce overfitting. The differing trends of the DGPs highlight the need for careful selection of `max_depth` and potentially favouring honest causal trees in scenarios characterized by higher noise or a bigger number of relevant covariates.



((a)) Max Depth vs. RMSE - Honest



((b)) Max Depth vs. RMSE - Adaptive



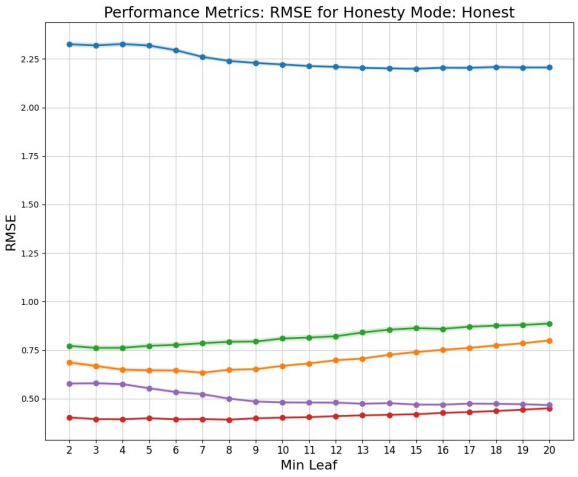
((c)) Max Depth vs. RMSE - T-Learner

Figure 1: Max Depth vs. MSE across multiple DGPs

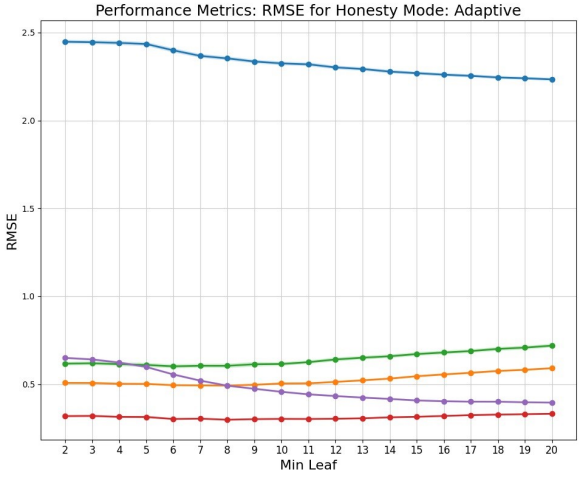
Minimum Leaf Size

Similarly, the influence of the minimum leaf size is dictated by the properties of the data, as illustrated in Figure 2. This hyperparameter acts as direct control over the granularity and interpretability of the learned tree structure. A distinct pattern is again noticed in Design 5 (Non-Linear CATE), Design 6 (Interaction Terms), and Design 7 with $\sigma = 0.1$ (Low Noise Level). For these DGPs, the MSE generally initiates at a lower level for smaller `min_samples_leaf` values, occasionally showing a slight dip, before progressively ascending as the value of the hyperparameter increases. This indicates that DGPs that feature complex CATEs or that are simple need finer-grained partitions to accurately approximate their underlying functions.

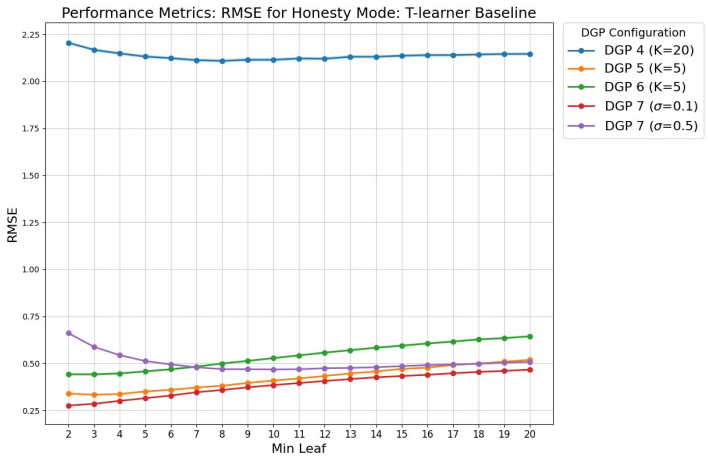
In comparison, Design 4 (Medium-Dimensionality with Relevant Covariates) and Design 7 with $\sigma = 0.5$ (Moderate Noise Level) showcase an opposing behaviour as the minimum leaf size is varied. For these scenarios, there is a continuous reduction in MSE, highlighting the need for having larger leaves to counter the effects of the noise and to enhance generalization. Thus, it becomes apparent that the complexity of the true CATE function is a factor to be taken into account when selecting the value of the `min_samples_leaf` hyperparameter.



((a)) Min Leaf vs. RMSE - Honest



((b)) Min Leaf vs. RMSE - Adaptive



((c)) Min Leaf vs. RMSE - T-Learner

Figure 2: Min Leaf vs. MSE across multiple DGPs

Overall, the previous plots show that the standard errors are low across all designs and metrics, as the shaded area is quite narrow for all three methods, which underscores the high precision and robustness of the simulation results. This allows for the interpretation of the observed differences in mean performance with confidence, as they are highly unlikely to be artifacts of simulation-specific random variation.

4.1.2 Comparison of Models

Table 2 presents the mean MSE, Bias^2, and Variance across all DGPs and models, with the standard errors available in Appendix E.

A notable finding across both hyperparameter configurations is that for Design 4, representing medium dimensionality relevant covariates, the overall MSE is among the highest, namely ≈ 5, while for the other designs the MSE is comparatively lower, less than 1. Yet, for Design 4, the honest causal tree consistently

achieves a lower MSE than the adaptive version (5.3761 vs. 5.9282 and 4.9340 vs. 5.4043). This demonstrates honesty's advantage in variance control (1.6314/0.9257) compared to the adaptive tree (3.3479/2.5026), despite a slightly higher bias. However, the T-Learner baseline for Design 4 achieves the lowest MSE (4.9765/4.4698), suggesting that a simpler approach can be more robust in this specific scenario.

In contrast, for Design 5 and Design 6, a different behaviour is noticed. Here, the adaptive causal tree consistently outperforms the honest causal tree in terms of MSE (Design 5: 0.4169 vs. 0.2516 and 0.4469 vs. 0.2541; Design 6: 0.5958 vs. 0.3716 and 0.6551 vs. 0.3783). This is justified by the significantly lower bias that the adaptive causal trees achieve (Design 5: 0.0110 vs. 0.0041 and 0.0129 vs. 0.0043; Design 6: 0.0105 vs. 0.0039 and 0.0141 vs. 0.0040), indicating that using the full dataset for splitting is critical for modelling complex non-linear relationships and interaction effects. The T-Learner baseline, however, achieves the lowest MSE in both Designs 5 (0.1196 and 0.1671) and Design 6 (0.2058 and 0.2788), primarily due to its low bias, suggesting that for these CATEs, a robust estimation of the conditional mean outcomes can lead to better CATE estimation.

The impact of noise level is particularly evident when comparing Design 7 with different values for the standard deviation of the error term. For the low noise scenario, with $\sigma = 0.1$, the adaptive causal tree achieves the lowest MSE (0.0983 and 0.0916), due to very low bias (0.0024 and 0.0021) and moderate variance (0.2697 and 0.2306), indicating accurate treatment effect identification without significant overfitting in clean data. Moreover, the T-learner baseline performs well in this case as well, achieving a competitive MSE (0.0780 and 0.1479).

However, as the noise level increases to $\sigma = 0.5$, the honest tree achieves a lower MSE (0.3059) for the maximum depth experiment compared to the adaptive tree (0.3574). This advantage stemmed from the honesty's variance control (0.4149 vs.0.5406) under moderate noise, even with slightly higher bias. This highlights how honesty's data separation acts as a beneficial regularization, preventing the tree from learning noise and improving generalization. The T-Learner baseline shows a higher MSE (0.5212) in this moderate noise setting, suggesting that it becomes less robust when the noise is increased.

Table 2: Mean values for MSE, Bias^2, Variance (Maximum Depth/Minimum Leaf Size)

DGP Scenario	**Metric**	**Honest Causal Tree**	**Adaptive Causal Tree**	**T-Learner Baseline**
Design 4 ($K = 20$)	MSE	5.3761/4.9340	5.9282/5.4043	4.9765/4.4698
	Bias^2	0.0662/0.0599	0.0498/0.0445	0.0308/0.0243
	Var	1.6314/0.9257	3.3479/2.5026	3.2902/2.1637
Design 5 ($K = 5$)	MSE	0.4169/0.4469	0.2516/0.2541	0.1196/0.1671
	Bias^2	0.0110/0.0129	0.0041/0.0043	0.0005/0.0009
	Var	1.1101/0.9268	1.1714/1.0921	1.3412/1.3248
Design 6 ($K = 5$)	MSE	0.5958/0.6551	0.3716/0.3783	0.2058/0.2788
	Bias^2	0.0105/0.0141	0.0039/0.0040	0.0017/0.0025
	Var	0.7322/0.5241	0.8938/0.7651	0.9131/0.8010
Design 7 ($\sigma = 0.1$)	MSE	0.1591/0.1616	0.0983/0.0916	0.0780/0.1479
	Bias^2	0.0050/0.0074	0.0024/0.0021	0.0004/0.0010
	Var	0.2683/0.2189	0.2697/0.2306	0.3171/0.3839
Design 7 ($\sigma = 0.5$)	MSE	0.3059/0.2306	0.3574/0.2086	0.5212/0.2187
	Bias^2	0.0155/0.0146	0.0069/0.0057	0.0044/0.0027
	Var	0.4149/0.2828	0.5406/0.3719	0.7645/0.4458

Overall, the comparative performance of honest and adaptive causal trees and their efficacy relative to the T-Learner baseline are strongly tied to the specific characteristics of the data. For high dimensionality or potentially strong noise influence, methods that prioritize variance controls, like honest trees, often yield lower MSE. For these conditions, optimal settings tend to be smaller `max_depth` values and larger `min_samples_leaf` values. Conversely, for DGPs with complex or low-noise CATEs, the superior bias reduction offered by the adaptive causal trees often has overall better performance. Optimal hyperparameter settings in these cases involve larger `max_depth` values and smaller `min_samples_leaf` values.

4.2 Real-World Data

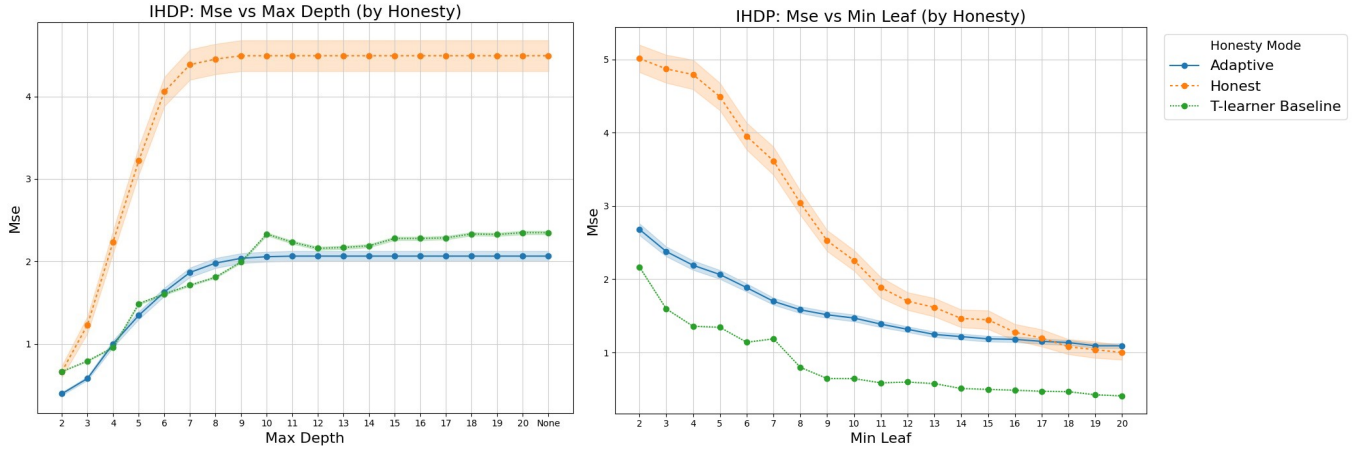
While the synthetic data experiments provided insights into the behavior of causal estimators under varying levels of complexity, it is essential to assess whether these findings generalize to real-world settings. Thus, the performance of causal trees is evaluated on the IHDP dataset by isolating the effects of the hyperparameters in subsubsection 4.2.1, but also studying how they interact with each other in subsubsection 4.2.2.

4.2.1 Isolating Hyperparameter Effects

To isolate the effects of each hyperparameter on the IHDP dataset, a single-parameter variation experiment was conducted, similar to the simulated data analysis.

Figure 3(a) showcases that as `max_depth` increases, the MSE increases rapidly before plateauing, indicating that deeper trees tend to overfit to spurious heterogeneity in the training data. This effect is particularly pronounced for honest causal trees, where the sample-splitting strategy reduces the observations available for leaf estimations, leading to unstable CATE estimates in deeper trees. This behaviour mirrors the patterns observed in the simulated Design 4 and Design 7, where increased model complexity, through dimensionality or noise, similarly led to a rise in MSE.

Similarly, Figure 3(b) highlights that increasing the `min_samples_leaf` value consistently lowers the MSE across all the models. This suggests that more aggregated leaves offer improved estimation of the CATE, aligning with observations from simulated designs involving dimensionality and noise, where coarser partitions mitigated the variance-induced overfitting.



((a)) MSE vx. Max Depth

((b)) MSE vs. Min Leaf

Figure 3: Effect of Hyperparameters on IHDP Dataset

As an overall observation across both plots, the honest causal tree consistently yielded the highest MSE with the largest standard error on the IHDP dataset. This result stems from the honesty constraint, which can severely limit the detection of meaningful treatment effect heterogeneity in moderately sized datasets like the IHDP, which has 747 observations, leading to high variance and suboptimal estimates. In contrast, the T-Learner baseline typically achieved the lowest MSE and standard error, likely because it fits two independent decision trees, benefiting from the full dataset size for training and yielding more stable predictions.

4.2.2 Grid Search over Both Hyperparameters

To explore the joint influence of `max_depth` and `min_samples_leaf` on model performance, a grid search was conducted. Figure 4 presents the corresponding heatmap of the Honest Causal Tree's MSE across these combined hyperparameter configurations.

The lowest MSE (0.6790) was achieved when both hyperparameters were set to 2, indicating that shallow trees with small leaves yield the most accurate CATE estimates on the IHDP dataset. A notable finding is the rapid deterioration in performance with even a slight increase in either hyperparameter, emphasizing the sensitivity of honest causal trees to tree complexity in real-world settings. This trend aligns with the isolated hyperparameter findings and is consistent with the theoretical properties of the honest framework, where sample splitting inherently limits overfitting in low-complexity trees.

While increasing the `min_samples_leaf` generally reduced MSE, suggesting that coarser tree partitions improve generalization, the effect is dominated by the tree depth. Deep trees consistently presented higher MSE across all configurations, suggesting that complexity-induced variance outweighs any bias reduction. These results reinforce the bias-variance trade-off observed in the simulation studies, providing practical guidance, namely that a small tree depth and a moderate leaf size achieve optimal balance for minimizing estimation error in honest causal trees on datasets like IHDP.

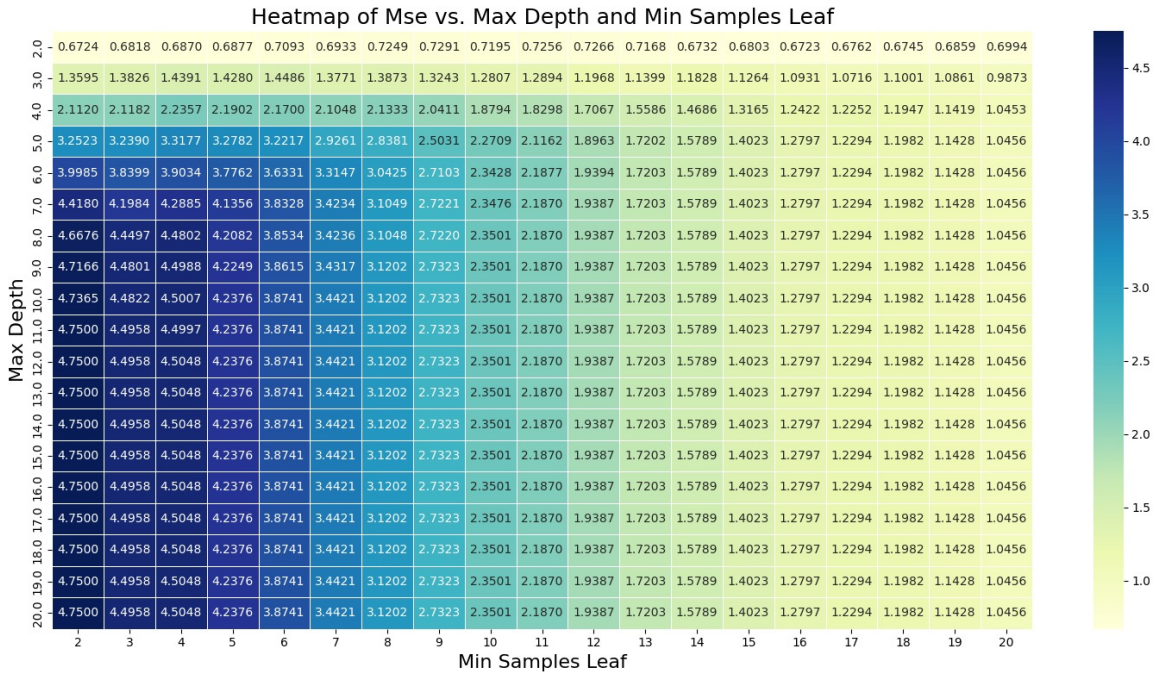


Figure 4: Effect of Min Leaf and Max Depth Combined on IHDP Dataset

In short, the analysis of the IHDP dataset revealed that increasing the `max_depth` generally increased the MSE for honest causal trees, while increasing the `min_samples_leaf` reduced it. Honest causal trees consistently showed higher MSE than the T-Learner baseline on this moderately sized dataset, suggesting sample-splitting limitations.

5 Discussion and Limitations

This section aims to interpret the key findings and to outline recommendations for addressing the study's limitations. In subsection 5.1, an analysis of the results presented in the previous section is provided. Following this, subsection 5.2 underlines the limitations of this study, while subsection 5.3 gives suggestions for how the current study can be enhanced with respect to the mentioned limitations.

5.1 Analysis of Results

The previous findings deepen the understanding of hyperparameter choice for honest and adaptive causal trees in CATE estimation. The results confirm that optimal hyperparameter choices for maximum depth and minimum leaf size are related to the underlying data characteristics. This aligns with the observation brought up by Hou and Fernández-Loría that the decision between adaptive and honest causal trees requires practical observation [7]. A comprehensive overview of the optimal hyperparameter settings can be found in Table 3.

Honest Causal Trees

On one side, honest causal trees demonstrated an advantage in high-dimensional and noisy environments, specific to Design 4 and Design 7 with $\sigma > 0.5$, due to enhanced variance control. This aligns with Athey and Imbens, whose sample-splitting strategy acts as a regularization, mitigating selection bias and preventing overfitting to noise [3]. This observation is also enforced by Hous and Fernández-Loría, who find that honesty particularly benefits models when the noise increases, confirming that variance reduction outweighs the loss of data efficiency in such cases [7]. Optimal settings in such data contexts involved smaller `max_depth` values and higher `min_samples_leaf` values for coarser, more robust partitions.

The expected U-shaped curve in the MSE plot, which supports the bias-variance trade-off, was prominent only in high-noise scenarios, signaling overfitting. Its frequent absence elsewhere suggests that causal trees offer robust control against severe overfitting.

Adaptive Causal Trees and T-Learner Baseline

In contrast, for data with complex structure (Designs 5 and 6) or low-noise settings (Design 7 with $\sigma = 0.1$), adaptive causal trees often outperformed honest causal trees due to significantly lower bias. Their ability to use the full dataset for both tree structure and effect estimation allows for capturing intricate relationships

[7]. It was found that the optimal hyperparameter settings involved larger `max_depth` values and smaller `min_samples_leaf` values for finer-grained partitions. Interestingly, the MSE curves for these data settings did not exhibit the hypothesized U-shaped curve. Instead, the MSE either plateaued or decreased, suggesting that increasing the model complexity through hyperparameters did not substantially increase variance. One plausible explanation is that in low-noise or highly structured data, the signal remains strong enough that deeper models do not overfit to random fluctuations.

An unexpected finding was the T-Learner's strong performance in certain complex CATE structures and on the real-world IHDP dataset, challenging the initial hypothesis that specialized causal tree algorithms would always yield better results. This observation is supported by Knaus et al., who showed that simple methods can sometimes outperform estimators in high-dimensional or small-sample settings [14]. The T-Learner's performance can be attributed to the fact that it fits two independent trees and computes the CATE as their difference. This approach can be more robust for non-linear or interactive CATEs, as it focuses on accurately estimating conditional mean outcomes. This suggests that causal trees are limited to settings that involve piecewise linearity.

Table 3: Hyperparameter Selection Guidelines for Causal Trees by Data Characteristic

Data Characteristic	**Recommended Model**	**Recommended Hyperparameter Values**
High-Dimensionality	Honest Causal Tree	Smaller `max_depth`, larger `min_samples_leaf`
Increasing Noise Levels		
Non-Linear CATE Structures	Adaptive Causal Tree or T-Learner Baseline	Larger `max_depth`, smaller `min_samples_leaf`
Interaction Terms		
Low-Noise Settings		
Small to Moderately Sized Real-World Datasets	Adaptive Causal Tree or T-Learner Baseline	Smaller `max_depth`, Larger `min_samples_leaf`

5.2 Limitations of the Current Study

This study is subject to several limitations, specifically the fact that it relies on standard causal inference assumptions, SUTVA, positivity, and unconfoundedness, which are rarely perfectly met in real-world observational data. While the IHDP dataset may reflect real-world confounding due to its retained original treatment assignment, all simulated data in this study were generated under the assumption of randomized treatment assignment, thereby excluding observed confounding variables. Furthermore, due to the use of a single tree (`n_estimators=1`) with the EconML library's `CausalForest` implementation, coverage could not be assessed, limiting insights into inferential guarantees, despite Athey and Imbens designing honest causal trees for valid inference [3].

Moreover, the explored hyperparameter range was systematic, but not exhaustive, and other influential hyperparameters were not varied. The simulated data provided simplified representations, so external validity to unmodeled real-world complexities should be interpreted with caution. Lastly, computational constraints limited the number of simulations and the dataset size.

5.3 Future Research Directions

Future research should focus on exploring other critical hyperparameters and employing more extensive optimization techniques on larger, diverse real-world datasets. Additionally, investigating robustness to violations of causal assumptions, including the presence and impact of confounding, would further enhance the practical applicability of these models. Finally, the study of coverage should be considered, possibly with a different causal tree implementation, in order to complement the accuracy-focused evaluations with measures of uncertainty and provide deeper insights into the estimator's reliability.

6 Responsible Research

This section reflects on the aspects taken into account for conducting responsible and reproducible research. The ethical implications of this research are discussed in subsection 6.1, followed by subsection 6.2, which analyses how reproducible the methods used in this study are. Lastly, subsection 6.3 describes how Large Language Models have been used responsibly.

6.1 Ethical Implications

Conducting research involving causal inference models necessitates careful consideration of ethical implications and adherence to responsible research practices. A primary concern in the application of CATE estimates, particularly in fields with direct human impact, is the potential for perpetuating existing biases or introducing unfairness. Models trained on historical data, even semi-synthetic datasets like the IHDP, may reflect societal biases or disparities present in the original observational data. If such biases are not addressed, the deployment of these models could lead to unfair recommendations or outcomes for different subgroups, such as misallocating resources or promoting discriminatory policies. Therefore, an ethical obligation involves inspecting the data for biases or employing strategies to mitigate their impact before applying the models to real-world applications. This includes, but is not limited to, fairness-aware machine learning techniques, re-weighing schemes, or disaggregated analysis of CATEs across different demographic subgroups to identify and address potential disparities. Furthermore, by providing guidelines for hyperparameter selection, this research contributes to the development of more accurate and reliable CATE estimates, which can enable more effective interventions.

6.2 Reproducibility and Transparency

Additionally, the reproducibility and transparency of the research methods are crucial for scientific integrity. To facilitate verification of the results presented in this study, the code developed and the IHDP dataset are publicly available [12]. All code is managed with version control (Git), and a `requirements.txt` file is provided to specify the exact software dependencies and versions used, ensuring the computational environment can be recreated. Moreover, a detailed documentation of the methodology has been provided, outlining the experimental steps, including the specific random seeds used for simulations and data splitting, to ensure full reproducibility of the numerical results. For all simulations, a fixed random seed of 42 was employed to ensure consistent and reproducible results.

6.3 Use of LLMs

Large Language Models, specifically ChatGPT and Gemini, have been used occasionally to assist in the writing and coding process. In terms of writing, LLMs have been used to ensure the flow of the paragraphs, for formatting tables and figures in LaTeX, and for correcting grammar errors. Additionally, they have been used for debugging purposes related to Python packages, improving computational performance, and plotting of graphs. It is important to note that LLMs served as a tool for writing and coding assistance, so the content generated was thoroughly reviewed and verified to ensure accuracy, originality, and alignment with the research's scientific integrity.

7 Conclusion

This study provides an empirical analysis of how key hyperparameters, maximum tree depth and minimum leaf size, influence the accuracy and generalization of CATE estimates in the context of causal trees, using both simulated and real-world datasets. The contribution of this research answers the central research question, demonstrating that optimal hyperparameter configurations are critically dependent on the underlying characteristics of the data, such as dimensionality, noise levels, and inherent complexity of the true CATE function.

The research analysed the behaviour of the two hyperparameters in multiple simulated data environments, as well as on the real-world IHDP dataset. A single-parameter variation was employed, and performance was assessed using the mean MSE, Bias^2, and Variance across multiple simulations.

The results established practical guidelines for hyperparameter selection, emphasizing the necessity of context-aware tuning in order to achieve a precise and generalizable causal effect estimation. Specifically, honest causal trees excel in environments with high noise, a large number of relevant covariates, or for large datasets, often performing best with shallow trees and large leaf nodes. In contrast, adaptive causal trees should be chosen for settings with low-noise, non-linear, or interactive CATES, or small to moderate-sized datasets, where deeper trees with smaller nodes can capture intricate relationships better. While optimal choices are linked to the data characteristics, with honest causal trees favouring variance control and adaptive causal trees or T-Learners performing better in bias reduction for complex or low-noise data, the outperformance of honest causal trees was sometimes unexpected, revealing their limitations due to sample splitting in data-constrained environments.

Building upon these findings, future research can explore several directions. These should involve validation on entirely real-world observational datasets, investigating the robustness of these models to violations of core causal assumptions, exploring alternative causal tree implementations that allow for the assessment of inferential guarantees, and the development of more sophisticated CATE estimates specific for real-world applications. By

providing an empirical roadmap for hyperparameter selection, this study helps in understanding the practical application of causal decision trees.

A Simulation on the Athey and Imbens DGPS (Designs 1, 2, and 3)

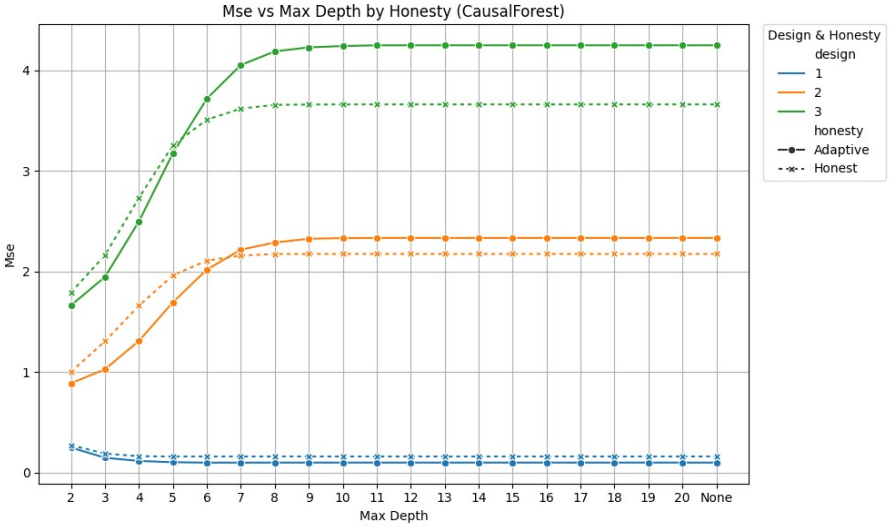


Figure 5: Athey and Imbens: Max Depth vs. MSE

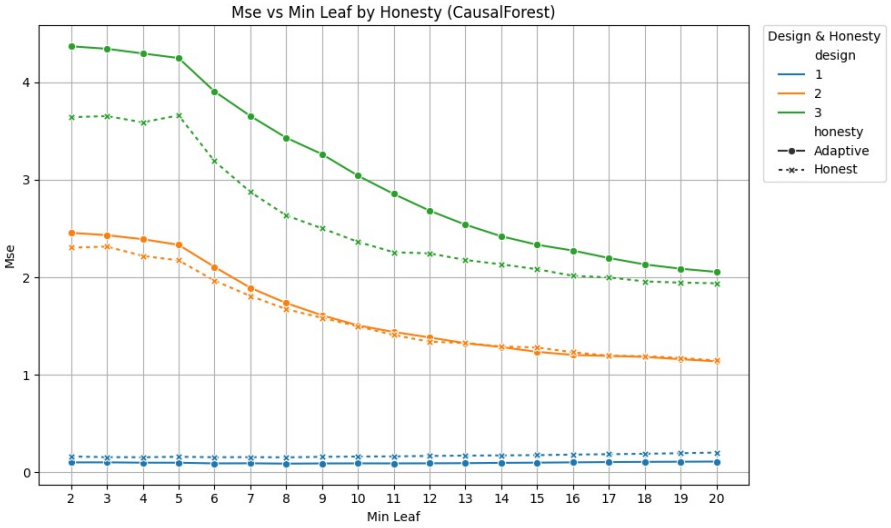


Figure 6: Athey and Imbens: Min Leaf vs. MSE

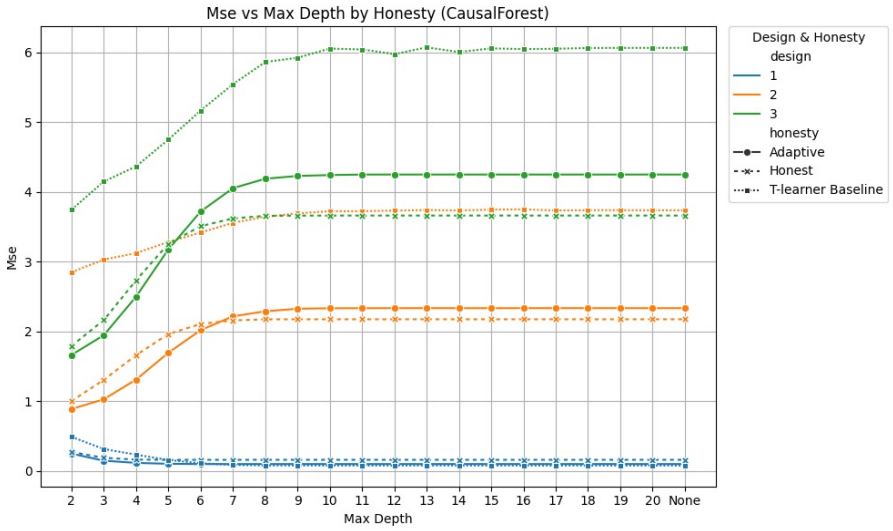


Figure 7: Athey and Imbens: Max Depth vs. MSE - T-Learner

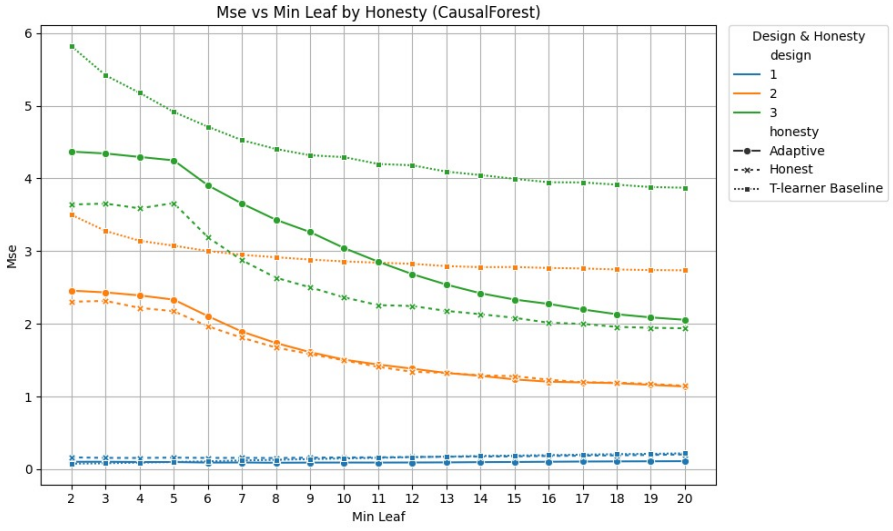


Figure 8: Athey and Imbens: Min Leaf vs. MSE - T-Learner

B Impact of Varying the Number of Relevant Covariates (Design 4)

The study of understanding how CATE estimation scales with increasing dimensionality when all covariates are relevant to the treatment effect was motivated by preliminary observations regarding Design 1. In Athey & Imbens (2016) [3], it is mentioned that Design 1 exhibits a comparatively smaller overall MSE for adaptive causal trees than honest causal trees, as observed in Appendix A, due to the sample splitting, where the benefit of bias reduction in such a simple DGP is outweighed by the loss of precision from using a smaller sample for estimation. Thus, it can be hypothesized that in more complex scenarios, where the number of relevant covariates is increased, the benefits of honesty will become apparent. To test this, the number of relevant covariates (K) in Design 4 was varied, using the values 2, 5, 10, 15, 20, 25, and 50.

The experiment confirmed that increasing dimensionality generally leads to a higher MSE across all models, indicating that accurately estimating CATE in higher dimensions is inherently more challenging. However, as hypothesized, honest causal trees demonstrated improved performance relative to adaptive trees when the number of covariates was sufficiently high. The T-Learner baseline achieved a lower MSE rather than the causal trees in low to moderate complexity settings, only being outperformed by honest causal trees at higher complexities.

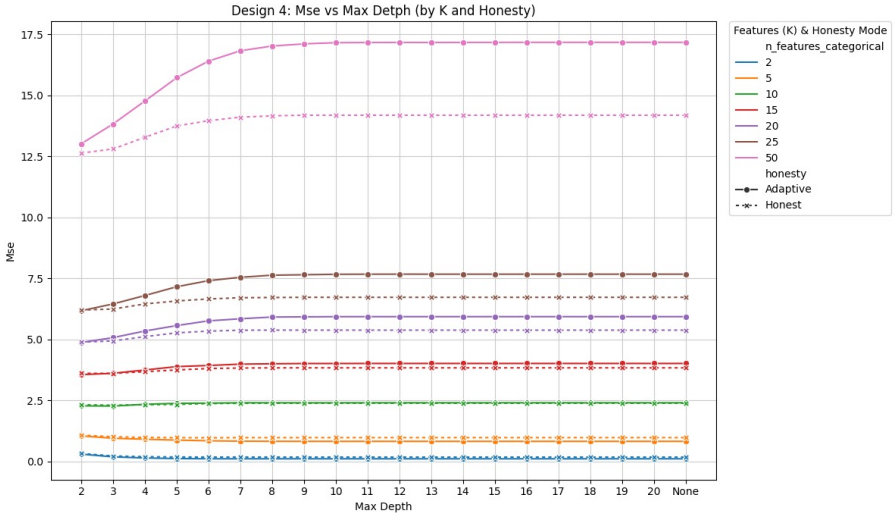


Figure 9: Design 4: Max Depth vs. MSE

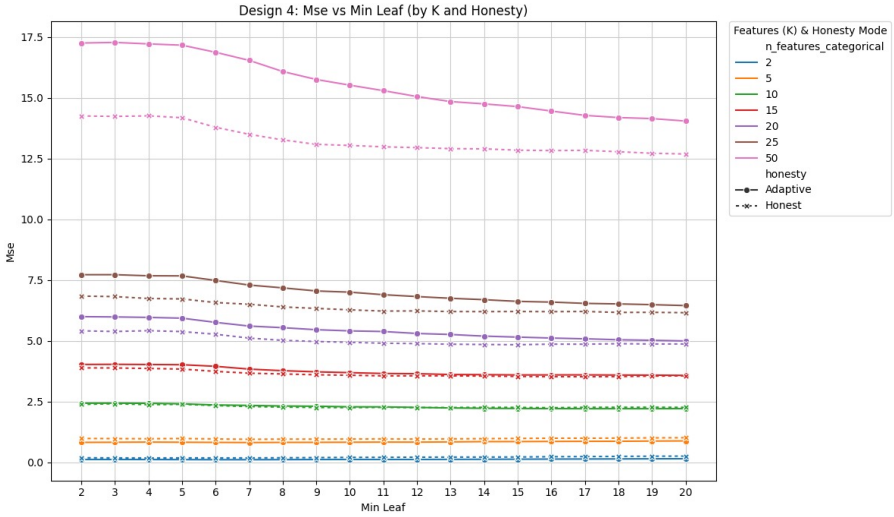


Figure 10: Design 4: Min Leaf vs. MSE

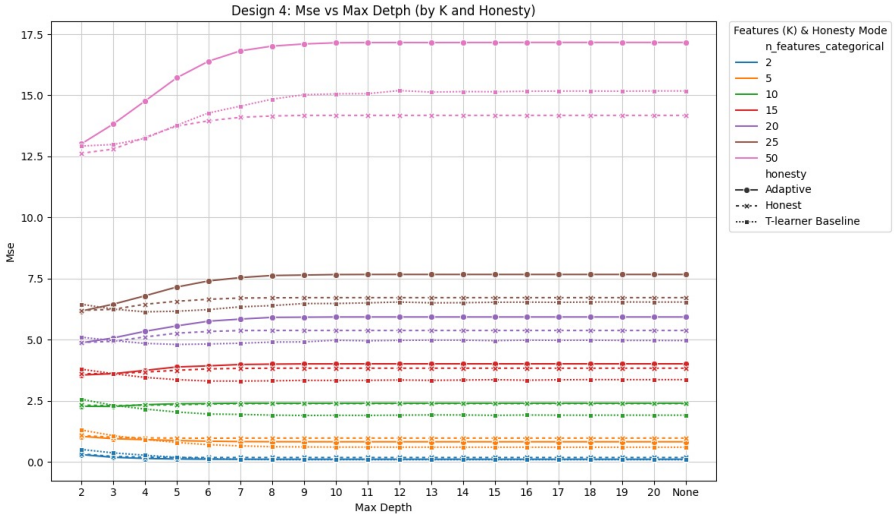


Figure 11: Design 4: Max Depth vs. MSE - T-Learner

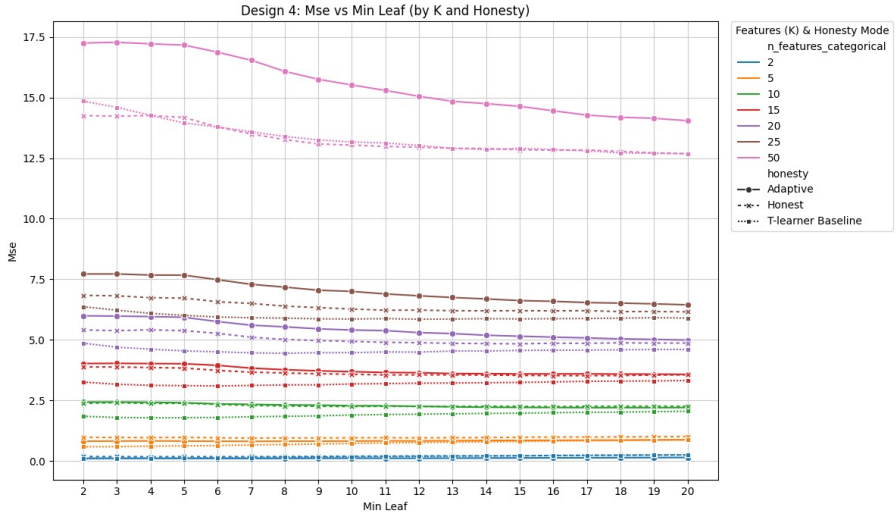


Figure 12: Design 4: Min Leaf vs. MSE - T-Learner

C Impact of Different CATE Structures (Designs 5, 6)

This experiment was specifically designed to isolate the effects of underlying functional forms of the CATE beyond piecewise linearity. To investigate the effects of non-linear and interaction-term CATE structures on causal trees, Designs 5 and 6 were used, both with low dimensionality (K=2, 5) and low noise($\sigma = 0.1$).

As opposed to the other designs studied, the shape of the curve of `max_depth` first decreases, after which it plateaus. The same contrastive behaviour is observed also for the `min_samples_leaf`, suggesting that introducing non-linearity and interaction terms affects the causal trees in different ways than the piecewise linearity of other DGPs. Moreover, varying the number of features K did not significantly affect the MSE values, implying that this parameter of the design does not play an important role in the CATE estimation. An interesting behaviour observed is the fact that the honest causal tree always has the highest MSE, highlighting the cost of honesty, while the T-Learner baseline achieved the lowest MSE among all models, suggesting its inherent base decision tree structure is well-suited for approximations for non-linear and interaction terms CATE structures.

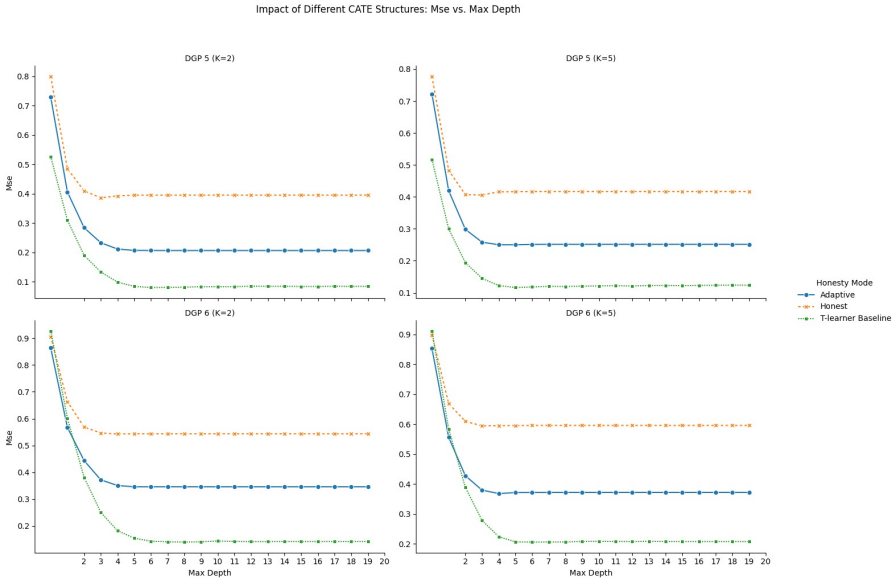


Figure 13: Design 5 and 6: Max Depth vs. MSE - T-Learner

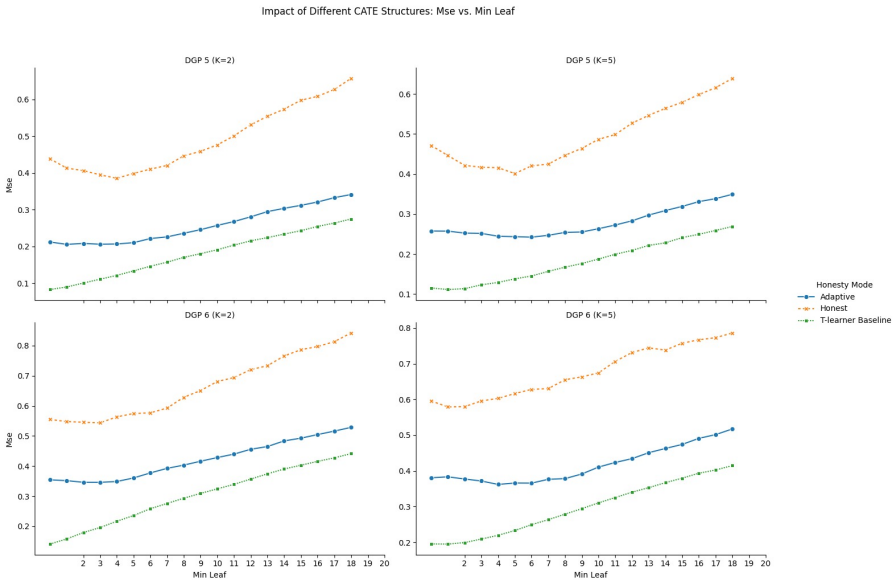


Figure 14: Design 5 and 6: Min Leaf vs. MSE - T-Learner

D Impact of Varying the Noise Level (Design 7)

Although Athey and Imbens propose different DGPs with different levels of complexity in terms of noise covariates, there is a lack of studying the effects of different levels of noise. To assess the robustness of honest and adaptive causal trees under systematically different levels of outcome uncertainty, thereby simulating varying real-world data quality, an experiment has been conducted in which the standard deviation of the error term varies. Moreover, following observations from the three Athey and Imbens DGPs, the MSE curve takes different shapes for different design complexities, thus this experiment aims to investigate whether increasing the noise level would induce a change in this curve's shape, and at what point such a change might occur. This has been done using Design 7, with K=2 and $\sigma \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1\}$, in order to reveal trends in the estimator's performance. A general expectation is that by increasing the noise level, the causal tree's ability to correctly estimate the CATE will decrease for both honest and adaptive trees.

As expected, the overall MSE increases as the level of noise increases across all estimator types for both hyperparameters, confirming that accurate CATE estimation becomes more challenging as data quality degrades. A key finding was the transition in the shape of the `max_depth` MSE curve, specifically the fact that while it stabilises for low noise, it tended towards a U-shaped curve at higher noise levels, confirming that noise can indeed induce classical overfitting behaviour. On top of this, it was noted that honest causal trees consistently outperformed adaptive causal trees in high-noise environments, highlighting an advantage of the honest approach. The T-Learner baseline model generally performs worse than honest causal trees in both cases, across most of the model hyperparameter configurations. This suggests that a specialised tree-based causal estimator offers an advantage in environments where noise is present.

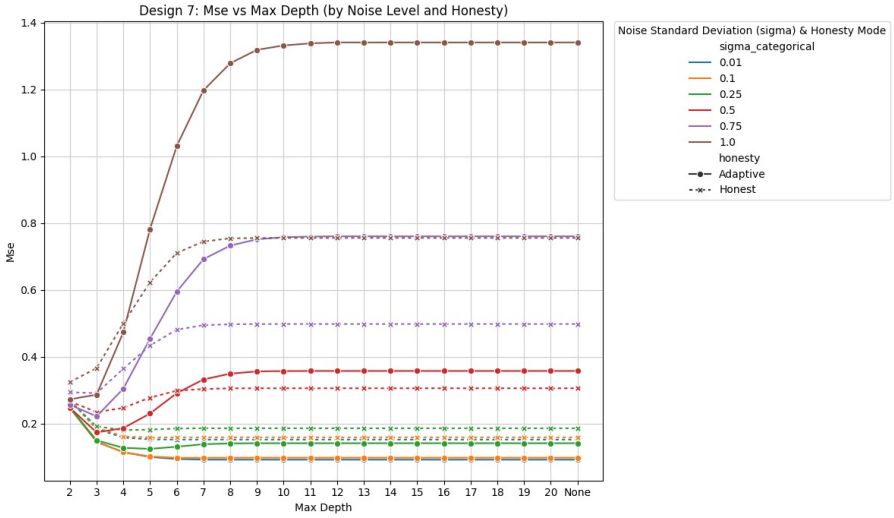


Figure 15: Design 7: Max Depth vs. MSE

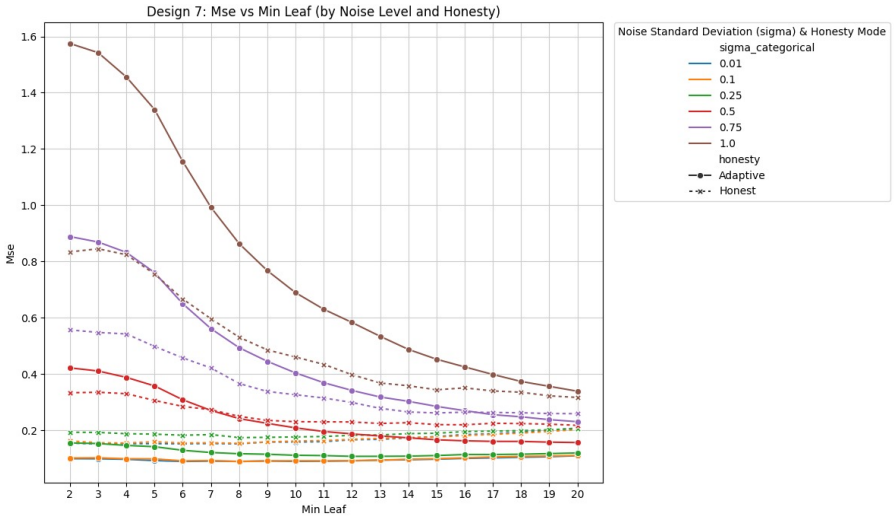


Figure 16: Design 7: Min Leaf vs. MSE

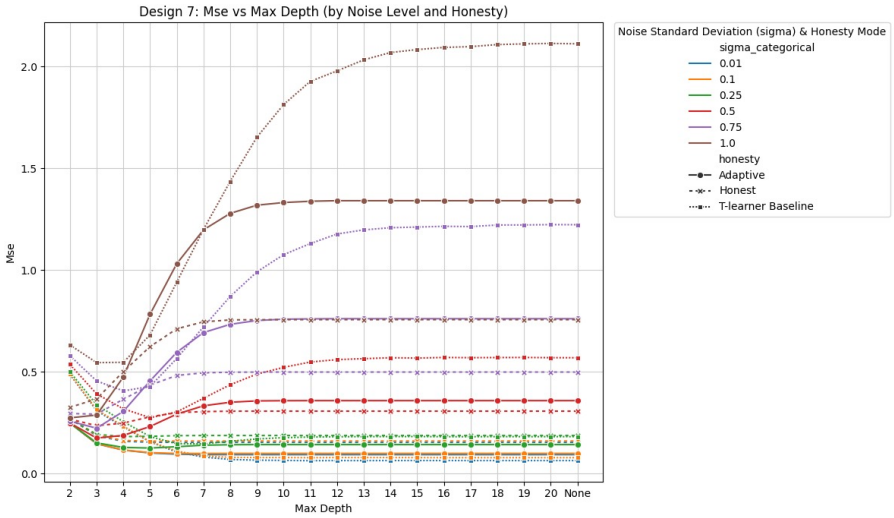


Figure 17: Design 7: Max Depth vs. MSE - T-Learner

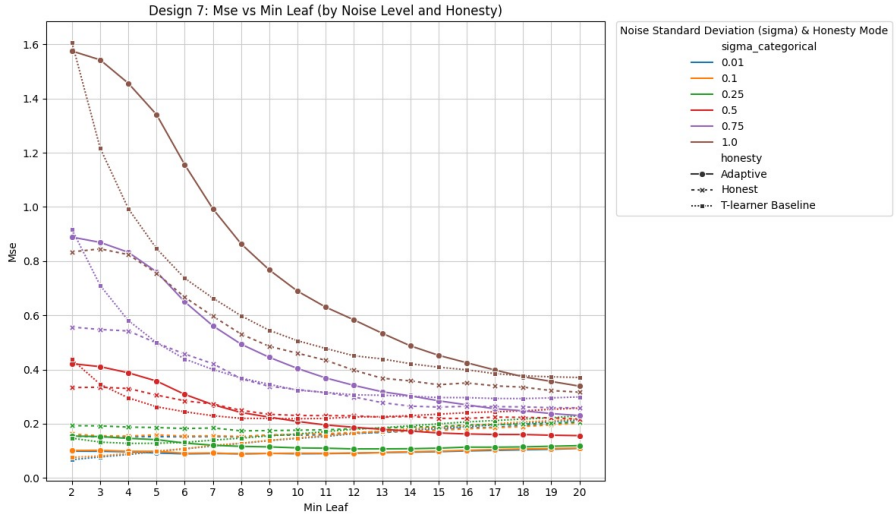


Figure 18: Design 7: Min Leaf vs. MSE - T-Learner

E Mean Values for Metrics for Standard Error

Table 4: Mean values for MSE, Bias^2, Variance (Maximum Depth/Minimum Leaf Size ± Standard Error)

DGP Scenario	Metric	Honest Causal Tree	Adaptive Causal Tree	T-Learner Baseline
Design 4 ($K = 20$)	MSE	5.3761 ± 0.0554/4.9340 ± 0.0429	5.9282 ± 0.0560/5.4043 ± 0.0494	4.9765 ± 0.0419/4.4698 ± 0.0426
	Bias^2	0.0662 ± 0.0077/0.0599 ± 0.0086	0.0498 ± 0.0064/0.0445 ± 0.0057	0.0308 ± 0.0038/0.0243 ± 0.0033
	Var	1.6314 ± 0.0642/0.9257 ± 0.0441	3.3479 ± 0.0578/2.5026 ± 0.0490	3.2902 ± 0.0424/2.1637 ± 0.0281
Design 5 ($K = 5$)	MSE	0.4169 ± 0.0135/0.4469 ± 0.0093	0.2516 ± 0.0062/0.2541 ± 0.0054	0.1196 ± 0.0022/0.1671 ± 0.0028
	Bias^2	0.0110 ± 0.0018/0.0129 ± 0.0020	0.0041 ± 0.0007/0.0043 ± 0.0006	0.0005 ± 0.0001/0.0009 ± 0.0001
	Var	1.1101 ± 0.0226/0.9268 ± 0.0192	1.1714 ± 0.0177/1.0921 ± 0.0169	1.3412 ± 0.0150/1.3248 ± 0.0132
Design 6 ($K = 5$)	MSE	0.5958 ± 0.0206/0.6551 ± 0.0212	0.3716 ± 0.0143/0.3783 ± 0.0128	0.2058 ± 0.0067/0.2788 ± 0.0068
	Bias^2	0.0105 ± 0.0016/0.0141 ± 0.0022	0.0039 ± 0.0006/0.0040 ± 0.0006	0.0017 ± 0.0002/0.0025 ± 0.0003
	Var	0.7322 ± 0.0255/0.5241 ± 0.0223	0.8938 ± 0.0206/0.7651 ± 0.0170	0.9131 ± 0.0135/0.8010 ± 0.0123
Design 7 ($\sigma = 0.1$)	MSE	0.1591 ± 0.0048/0.1616 ± 0.0045	0.0983 ± 0.0040/0.0916 ± 0.0026	0.0780 ± 0.0012/0.1479 ± 0.0021
	Bias^2	0.0050 ± 0.0007/0.0074 ± 0.0010	0.0024 ± 0.0003/0.0021 ± 0.0003	0.0004 ± 0.0001/0.0010 ± 0.0001
	Var	0.2683 ± 0.0077/0.2189 ± 0.0080	0.2697 ± 0.0061/0.2306 ± 0.0049	0.3171 ± 0.0035/0.3839 ± 0.0050
Design 7 ($\sigma = 0.5$)	MSE	0.3059 ± 0.0091/0.2306 ± 0.0073	0.3574 ± 0.0076/0.2086 ± 0.0052	0.5212 ± 0.0072/0.2187 ± 0.0036
	Bias^2	0.0155 ± 0.0019/0.0146 ± 0.0017	0.0069 ± 0.0009/0.0057 ± 0.0009	0.0044 ± 0.0007/0.0027 ± 0.0004
	Var	0.4149 ± 0.0139/0.2828 ± 0.0124	0.5406 ± 0.0110/0.3719 ± 0.0100	0.7645 ± 0.0097/0.4458 ± 0.0075

References

[1] Stefan Feuerriegel, David Frauen, Viktor Melnychuk, et al. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(5):958–968, 2024.

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

[3] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[4] Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):587–593, 1980.

[5] Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[6] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[7] Yanfang Hou and Carlos Fernández-Loría. Honesty in causal forests: When it helps and when it hurts. *arXiv:2506.13107*, 2025.

[8] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[9] Damian Machlanski, Spyridon Samothrakis, and Paul Clarke. Hyperparameter tuning and model evaluation in causal effect estimation. *arXiv:2303.01412*, 2023. Version 1, submitted 2 March 2023.

[10] Aapo Hyvärinen, Shohei Shimizu, and Patrik Hoyer. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity. pages 424–431, 2008.

[11] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Working Paper 24678, National Bureau of Economic Research, 2018.

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

[13] Microsoft Research. EconML: A python package for ml-based heterogeneous treatment effects estimation. `https://econml.azurewebsites.net/index.html`, 2019.

[14] Maximilian C. Knaus, Michael Lechner, and Alexander Strittmatter. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *arXiv:1810.13237*, 2018. 1st version October 31, 2018; revised December 18, 2018.

[15] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.