University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Data Science

Empirical Exploration of Explanation-Guided Learning for Causal Effect Estimation

Supervisor

Prof. Alberto Testolin
University of Padova

Co-supervisor

Prof. Sam Verboven
Vrije Universiteit Brussel

Master Candidate

Anna Glado
2122285

Academic Year

2025-2026

Contents

Abstract

This thesis investigates the integration of Explanation-Guided Learning (EGL) into causal inference, with a focus on Conditional Average Treatment Effect (CATE) estimation from observational data. While EGL has been widely studied in areas such as computer vision and natural language processing with an objective to improve explainability and generalizability of models, its application in the domain of causal inference remains largely unexplored. We propose a framework that incorporates EGL, specifically with counterfactual explanations, into modern CATE estimators: Meta-Learners. We improve accuracy of various meta-learning architectures by integrating explanation-based loss following the latest practices of the EGL literature. The proposed method enables the integration of structured domain knowledge into causal learning pipelines, which to the best of our knowledge, has not been done before. To evaluate the effectiveness of this approach, we conduct controlled experiments using synthetic data with known ground-truth treatment effects and varying levels of confounding. We assess performance across several meta-learners, including S-, X-, DR-, and domain adaptation variants. Our results demonstrate that incorporating explanation signals improves the accuracy of CATE estimation. In particular, the DA-learner and S-learner show the greatest reductions in mean absolute bias, with decreases of 49% and 51.8%, respectively. Overall, this work bridges two previously disconnected research areas – a newly emerged EGL and causal inference on observational data – and provides experimental proofs that demonstrate how explanation signals improve the accuracy of CATE estimation.

Acknowledgments

This thesis was written at the Data Lab of Vrije Universiteit Brussel (VUB) under the co-supervision of Professor Sam Verboven and PhD students Alessandro Marchese and Luc Hirsch. This collaboration was very much custom, spontaneous, and meaningful, exceeding many of my expectations.

Back in August, after I had just finished my internship at Microsoft in Brussels, I reached out to several well-known Belgian universities and their data science labs, gave a short elevator pitch about who I am and what I am curious about, and asked whether anyone had a project I could contribute to. I was not sure whether this was even possible from an official point of view, but I wanted to shoot my shot by being honest, curious, and proactive – and it worked. Professor Sam Verboven proposed that I join several ongoing projects at a research lab at VUB: running experiments for an ongoing research paper on flexible counterfactual explanations (Stig Hellemans et al.) and a project on EGL in causality, the latter becoming my main thesis topic.

Sam Verboven also made sure that I was well integrated into the lab, so that I never felt alone, which I am incredibly grateful for. I was surrounded by helpful, open-minded PhD students who quickly became my friends throughout my time at VUB. Because of them, this thesis project, although very challenging at times, felt like a lot of fun. I was invited to all the lab meet-ups and team-building activities and never felt like an impostor. Sam Verboven and all the lab members accepted me fully and made me feel at home. With their help and encouragement, I gained lots of confidence, which allowed me

to grow so much academically.

It was with this new mindset that I followed my professor's advice and presented at an ORBEL conference in February in Leuven. Demonstrating this work in front of researchers from all over Belgium felt like a major highlight of this project, and what was even more rewarding was how positive the response from the audience was.

I want to say a big thank you to my supervisor, Alberto Testolin, who has always been open-minded and supportive of my ideas, and to my co-supervisor, Sam Verboven, for believing in me and presenting me with this incredible opportunity. I also thank Alessandro Marchese for all the guidance, spontaneous syncs, coding help, and his ability to magically understand my messy questions and entangled thought processes. And of course, a big thank you to the other members of the lab — Luc Hirsch (for the crash course in causality), Mathias Hanson (he is too funny, so sitting next to him is the worst for productivity), Emilie Gregoire, Professor Alexander Thys (or simply, THE Alex — the soul of Data Lab in my opinion), and others. The VUB cafeteria also deserves a special thank you — it really wasn't the best, and yet, it brought us all together.

It has been such a wholesome and rewarding experience that I even started considering getting a PhD – stay tuned!

Declaration of usage of generative AI

AI technologies were used in this thesis solely for proofreading and identifying formatting issues. In the coding process, AI tools were occasionally employed to troubleshoot errors and resolve minor syntax issues. All content was carefully reviewed and edited by the author, who assumes full responsibility for the accuracy and integrity of the work.

Acronyms

ATE — Average Treatment Effect
$\tau = E[Y(1) - Y(0)]$
CATE — Conditional Average Treatment Effect
$\tau(x) = E[Y(1) - Y(0) \mid X = x]$
CF — Counterfactual (e.g., counterfactual explanations)
CV — Computer Vision
DA — Domain Adaptation (e.g., DA-learner)
DiCE — Diverse Counterfactual Explanations
DR — Doubly Robust (e.g., DR-learner)
EGL — Explanation-Guided Learning
ITE — Individual Treatment Effect
$\tau_i = Y_i(1) - Y_i(0)$
IML — Informed Machine Learning
ML — Machine Learning
NLP — Natural Language Processing
NN — Neural Network
PO — Potential Outcome
PW — Propensity-Weighted
RCT — Randomized Controlled Trial
VQA — Visual Question Answering
xAI — Explainable Artificial Intelligence

Chapter 1

Introduction

In many operational systems with cause-and-effect relationships, such as healthcare, retail, or finance, we are often interested in understanding whether a certain decision or intervention was effective. A modern approach to this problem is to leverage causal machine learning models, which are trained on data to learn patterns and estimate the impact of different factors on outcomes.

In these causal domains, human-generated explanations of outcomes are often naturally available [41]. In simple terms, these explanations describe why a certain decision was made or why a particular outcome occurred. They help us understand how different characteristics of a person, product, or situation influence decisions and outputs. These explanations can take many forms: reasons for refusing a request in resource allocation systems, explanations for why a product was returned, stated causes of customer churn, or structured feedback in recommender systems. For example, in the organ allocation domain, such explanations can appear as reasons for accepting or rejecting a transplant candidate – e.g., "the organ is rejected because the donor is too old" [41].

Importantly, these post-hoc explanations often encode valuable domain knowledge about the mechanisms underlying decisions and outcomes. How-

ever, despite their potential informational value, they are typically not used during model training – meaning that when models learn from data, these explanations are not included as part of the learning process and are instead discarded [57].

Recent advances in Explainable Artificial Intelligence (xAI) have focused on generating explanations to better understand the behavior of deep learning models. Unlike the human-provided explanations described above, these are produced by the model itself, but they serve a similar purpose: revealing how inputs influence outputs. The extension of xAI – Explanation-Guided Learning (EGL) – takes this further by proposing a paradigm of integrating human-provided explanations directly into the learning process — often in combination with model-generated explanations — as supervisory signals, guiding models toward more meaningful and robust representations [18].

At the same time, in the field of causal inference, one of the main quantities of interest is Conditional Average Treatment Effect (CATE), which aims to capture how causal effects vary across individuals with different characteristics. Estimating CATE is especially challenging in observational settings, where treatment assignment is not randomized. In such cases, estimation is vulnerable to systematic biases, most notably confounding, which can induce spurious correlations and lead to inaccurate CATE estimates [23].

These two research directions — EGL and CATE estimation — have largely evolved independently and might seem unrelated. However, we argue that there is significant potential to bridge them. This intersection is the core idea of this thesis, which brings us to the main hypothesis: *explanation signals can serve as a structured form of domain knowledge to improve CATE estimation.*

Our motivation is further supported by recent work in organ allocation [41], which demonstrates that incorporating human-provided explanation signals can improve treatment effect estimation in practice. This example is particularly compelling, as it highlights a setting where high-stakes decisions

rely on both data and expert reasoning, yet current machine learning approaches typically ignore the latter.

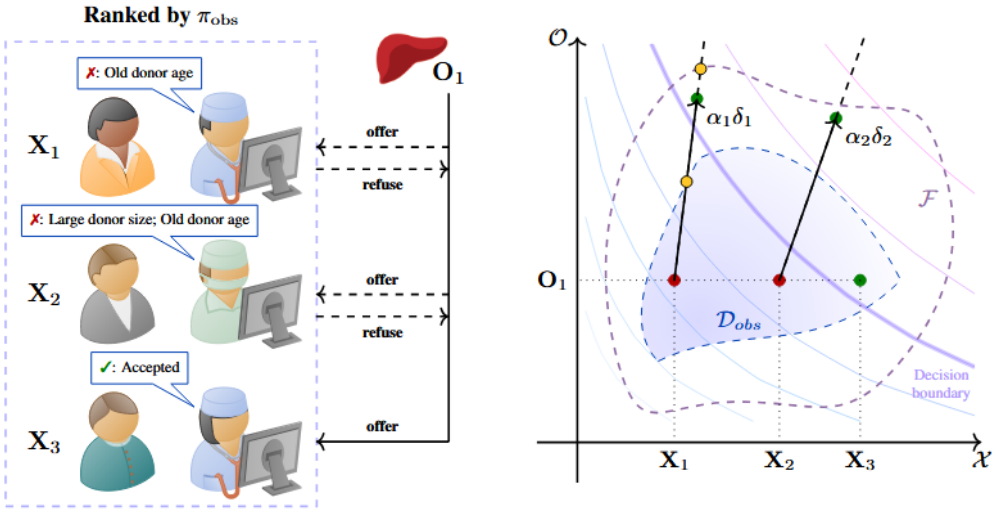

Figure 1.1: Example of explanation-guided learning in organ allocation: clinician-provided refusal reasons (e.g., "old donor age" or "organ is too big") are used as direction-only counterfactual signals to guide causal learning. The proposed model (ClexNet) leverages these signals to learn policy-invariant representations, resulting in improved predictive performance, generalization, and calibration compared to standard observational models [41].

This thesis begins by introducing the foundations of EGL, including its emergence, taxonomy, and mechanisms for integrating explanations into model training. We then turn to causality, focusing on CATE estimation in observational settings and the challenges posed by confounding and selection bias. Building on this foundation, we review modern approaches for CATE estimation, with particular emphasis on meta-learning frameworks and their neural network-based implementations, which form the methodological backbone of our experiments.

We then formalize the problem setting and introduce our methodological framework, representing it through a directed acyclic graph and corresponding structural equations. Based on this formulation, we propose a novel ob-

jective function that integrates EGL into causal estimation. Finally, we empirically evaluate the proposed approach using controlled experiments with synthetic data, assessing the impact of explanation signals on the accuracy of CATE estimation across different model architectures and configurations.

Despite the conceptual promise of combining EGL with causal inference, there is currently a lack of unified frameworks and systematic empirical studies that evaluate this integration. **Addressing this gap forms the core contribution of this thesis.**

Chapter 2

Literature Review

2.1 Emergence of Explanation-Guided Learning

The increasing complexity of modern artificial intelligence (AI) systems, together with regulatory initiatives aimed at transparency and accountability in automated decision-making (e.g., the *AI Act* [1]), has highlighted the critical need for interpretable AI. This has renewed attention to *Explainable Artificial Intelligence* (xAI), which focuses on understanding, interpreting, and communicating the reasoning processes of complex machine learning (ML) models.

At its core, xAI seeks to map model behavior—typically predictions or outputs—to interpretable components such as input features, training samples, or internal representations. A central concept underlying most xAI methods is *attribution*, which identifies the parts of the input or model responsible for a prediction [65].

EGL extends xAI beyond post-hoc analysis by incorporating explanation signals directly into the learning process. Instead of merely inspecting model behavior, EGL uses explanations to actively guide training, improving model

properties such as robustness, fairness, or alignment with domain knowledge. Formally, EGL is a learning paradigm in which explanations—provided by humans or generated by auxiliary models—are used as additional supervision during training [18, 50]. Instead of relying solely on predictive loss, EGL constrains the learning process using explanation signals via supervision, regularization, or data augmentation. That way, the trained model aligns with prior knowledge about which features should influence predictions.

2.1.1 EGL as a Special Case of Informed Machine Learning

From a broader perspective, EGL can be seen as part of a broader class of approaches that integrate external knowledge into machine learning systems. Such approaches are united by the paradigm of *Informed Machine Learning* (IML), which studies how prior knowledge can be systematically incorporated into the learning pipeline.

Indeed, the existence of IML paradigm demonstrates that guiding models with expert signals is not a novel idea. Von Rueden et al. [59] formalize this paradigm by proposing a comprehensive taxonomy of how prior knowledge can be integrated into learning systems. Their survey highlights a wide range of approaches, including labeling strategies, feature engineering, knowledge graphs, logical rules, algebraic equations, and even physical simulations. While this diversity underscores the richness of theory-guided data science, the heterogeneous nature of these approaches can make method selection non-trivial, particularly due to overlapping concepts and inconsistent terminology [59].

Within this framework, EGL can be interpreted as a special case of informed machine learning. According to the definition by von Rueden et al.—"learning from a hybrid information source that consists of data and prior knowledge, where the prior knowledge originates from an independent source, is formally represented, and is explicitly integrated into the learning

pipeline"—EGL satisfies these criteria when expert-provided explanations are treated as prior knowledge. In this setting, explanations such as feature attributions or counterfactual examples constitute structured signals that encode domain expertise about the relationship between inputs and outputs.

Focusing on integration at the level of the learning algorithm, the survey identifies several mechanisms for incorporating prior knowledge, including the use of informative priors as regularizers, structural constraints that guide model selection, qualitative constraints such as monotonic relationships, and causal constraints derived from domain ontologies. These approaches share the common principle of constraining the learning process to favor solutions consistent with prior knowledge.

Despite this breadth of methods, the survey does not explicitly consider the alignment between expert-provided knowledge and model-generated interpretability signals. This highlights the novelty of EGL within the informed machine learning framework: it introduces a mechanism for integrating prior knowledge by aligning expert explanations with model-generated explanations, thereby constraining not only the model's outputs but also **the way in which those outputs are derived**.

2.1.2 Taxonomy of EGL

A recent survey on EGL [18] categorizes its strategies along two complementary axes:

1. **Scope of guidance:** *Local* vs. *Global* explanations.
2. **Integration mechanism:** *Supervision*, *Regularization*, or *Data Augmentation*.

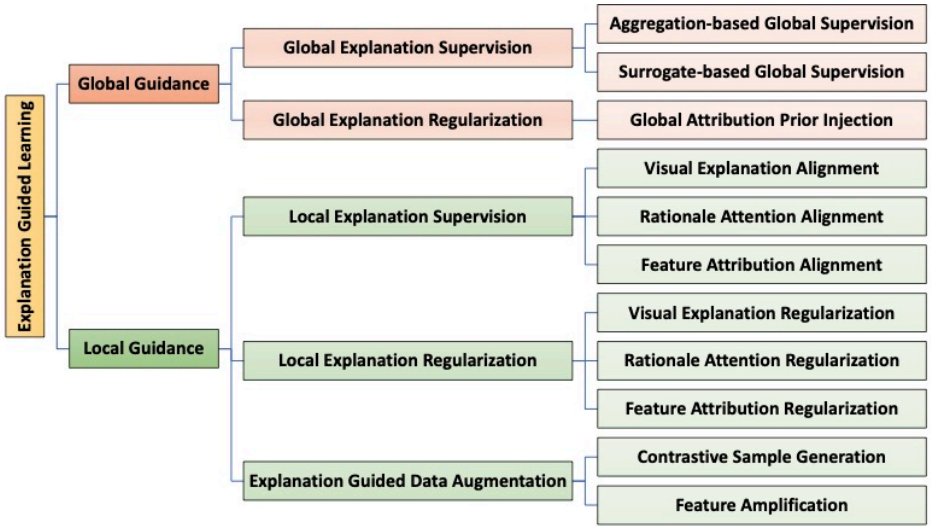

Figure 2.1: Taxonomy of EGL [18]

a) Local vs. Global Explanation Guidance

Local guidance operates at the level of individual data samples. Explanations are generated for each input, and supervision or regularization is applied per instance. Local approaches are flexible and can capture fine-grained patterns unique to each sample, making them widely used across domains [18].

Global guidance aims to generate a single, model-wide explanation, often by aggregating local explanations or training an interpretable surrogate model. This produces overall feature importance scores or interpretable global structures but may overlook localized behaviors.

Local and global approaches are not mutually exclusive; they can be combined with supervision, regularization, or data augmentation strategies to provide multi-scale guidance.

b) Integration Mechanisms: Supervision, Regularization, and Data Augmentation

Once the scope (local/global) is defined, explanations can be incorporated into learning through three primary mechanisms:

Explanation supervision assumes the availability of ground-truth explanation signals for each data sample. During training, these signals are compared against model-generated explanations. Ground-truth explanations may include annotated image regions, highlighted text tokens, or expert-provided rationales. A key challenge is the limited availability and heterogeneous format of such signals, particularly across domains like tabular data [44, 63].

Explanation regularization embeds prior knowledge about desirable explanation properties directly into the learning objective. For instance, models may be penalized for relying on spurious features or encouraged to focus on subsets of relevant features. Regularization is especially useful when ground-truth explanations are unavailable, as it enforces high-level constraints on model reasoning [43, 52].

Explanation-guided data augmentation leverages model explanations to generate additional training samples, aiming to improve generalization and robustness, especially in low-data or high-variance settings. Typical strategies include perturbing features identified as irrelevant, masking or modifying less important attributes, and constructing counterfactual instances that minimally alter predictions [26, 38]. These approaches can be applied in an unsupervised or weakly supervised manner, as they rely on model-derived signals rather than ground-truth annotations. Hasan et al. [26] provide a compelling case study demonstrating the effectiveness of counterfactual data augmentation on tabular datasets. Their experiments on the Adult dataset show that counterfactual-augmented training outperforms GAN-based augmentation, highlighting the potential of this approach for tabular data.

2.1.3 Types of Explanations

Before diving into the next section, it is important to reiterate the distinction between the explanations provided by experts and model-generated explanations. Expert explanations signify human/domain knowledge, while model-generated explanations are interpretability signals from a trained model. To avoid confusion, we refer to the latter as **model explanations.**

A model explanation is generated with an explainer, which is a model or an algorithm designed to explain or visualize model's decision-making process [18]. Formally, an explainer can be viewed as a function that takes as input a predictive model together with relevant data, such as feature inputs and model outputs. A wide variety of explainers exist, which explains why model explanations take numerous forms: saliency maps, feature importance vectors, rationales, counterfactual explanations, etc. [18]. Needless to say, such representations also vary depending on the domain of those data, e.g., visual, text-based, and tabular.

For example, for our context – tabular data – empirical observations by Borisov et al. [7] indicate that the most common forms of explanations are **attribution-based explanations** and **counterfactual explanations**. For both of these forms, EGL literature provides a systematic overview of commonly used explainers [18]. For feature attribution methods, the common choices are LIME [47] and SHAP [39], while for counterfactual explanation methods, a popular method is DiCE [45]. These approaches differ fundamentally in how they represent model behavior. Attribution-based methods assign importance scores to input features, providing insight into which variables drive predictions, while counterfactual methods describe how inputs must change to alter model outputs, offering a more intervention-oriented perspective.

A fundamental challenge when working with explainers is that different methods may produce conflicting explanations for the same model and input. These discrepancies can arise not only in the magnitude of feature attribu-

tions but also in their direction, i.e., whether a feature contributes positively or negatively to a prediction [36]. Such inconsistencies are particularly critical in EGL settings, where explanations serve as supervision signals. Misalignment between explainer outputs and expert-provided annotations can lead to substantially different optimization trajectories and ultimately affect the learned model behavior.

Counterfactual Explanations

In this work, we focus on counterfactual explanations, since they can be easily integrated into the learning pipeline through data augmentation [18]. Moreover, with the aforementioned strategy, we avoid the challenges of non-differentiable explainers such as LIME.

Counterfactual explanations identify minimal changes to an input that would lead to a different model prediction. Formally, they aim to solve an optimization problem that balances proximity to the original instance with achieving a desired target outcome [60]. These explanations are particularly intuitive, as they directly answer "what-if" questions and align closely with human reasoning about causality and decision-making.

Counterfactual explanations generated from expert knowledge can naturally complement EGL. These explanations can act as synthetic training samples that are integrated into the learning pipeline through EGL augmentation technique. By passing both a batch of original training data and a counterfactual batch to the model, we can combine the effects of the observational data and expert knowledge. Since these counterfactual samples are generated from expert knowledge, their distribution is expected to be more closely aligned with the ground truth, which means they have the potential to de-bias causal estimators.

We now transition to the second research domain, which also constitutes the primary context of our study and experiments: causal inference.

2.2 Causal Inference

2.2.1 Conditional Average Treatment Effect

As introduced earlier, this thesis aims to leverage EGL techniques in the area of causality, where a central question is how effective a particular intervention is. In other words, we are estimating the treatment effect. This quantity is of significant interest to various scientific communities, such as in medicine [17, 19] and social sciences [29] for assessing the efficacy of a policy. Several quantities are commonly used to measure causal effects, including the individualized treatment effect (ITE), the average treatment effect (ATE), and the CATE [28]. In this work, we focus on the **CATE**, which describes how the treatment effect varies depending on the characteristics of individuals, typically represented by observed covariates. Estimating CATE allows us to analyze treatment effect heterogeneity across a population. This makes it particularly useful in practical applications, as it helps to identify which subgroups benefit more—or less—from a given intervention. For example, CATE estimation can help determine which customer segments respond most positively to a marketing campaign or which groups of patients are more likely to recover after receiving a specific treatment.

Recently, various methods have been developed using machine learning to estimate CATE [3, 15, 22, 53, 56, 61]. Although these methods have proven successful, their effectiveness in estimating treatment effects can be significantly compromised in real-world applications that rely on observational data [5].

2.2.2 CATE on Observational Data

In the perfect scenario, we would like to estimate treatment effect on data generated under an RCT setup [51], which is a form of scientific experiment where allocation of subjects into treatment and control groups is fully

randomized. And because of such randomization, the effect of other non-causal factors is nullified. But in practice, many research questions cannot be studied using RCTs [23]. RCTs are often expensive and time-consuming, particularly when investigating rare outcomes. Moreover, they frequently involve highly selected populations, which limits the generalizability of their findings. In some cases, randomization is not feasible at all due to ethical or practical constraints [49]. As a result, observational data may be the only available resource for causal estimation.

In observational settings, treatment assignment is not randomized, so some subpopulations are more likely to receive treatment than others. In other words, the propensity — the probability of receiving treatment given a set of observed covariates — is not independent of these covariates as it is in an RCT, which makes isolating and estimating causal effects more challenging. Nevertheless, although the lack of randomization complicates causal inference, meaningful causal insights can still be obtained if appropriate assumptions hold and suitable statistical adjustments are applied [28].

Overall, estimating causal quantities from observational data presents challenges that extend beyond predictive modeling. Unlike randomized experiments, observational datasets reflect the combined effects of treatment assignment mechanisms, data collection processes, and temporal dynamics [23]. Consequently, observed associations may not correspond to the causal relationships of interest.

Identification Assumptions

A central question in observational causal inference is whether a target causal estimand is identifiable from the observed data. Identification analysis determines whether a causal quantity can be uniquely recovered under a set of assumptions, or only bounded in partially identifiable settings [6, 40].

In observational studies, identification commonly relies on the assumption of *conditional ignorability* – no unmeasured confounding – which also means treatment assignment is independent of potential outcomes given observed

covariates [49]. Under this assumption, observational data can be interpreted as arising from a conditionally randomized experiment if the following conditions hold [28]:

- *Consistency*: Observed treatments correspond to well-defined interventions.
- *Conditional exchangeability*: Treatment assignment is independent of potential outcomes given covariates L.
- *Positivity*: Each treatment level has positive probability for all L.

However, in practice, these assumptions are fundamentally untestable from the data alone and rarely fully satisfied in practice, making causal conclusions dependent on modeling choices and domain knowledge [21]. In particular, the assumption of unobserved confounders – which cannot be empirically verified – can violate exchangeability and introduce substantial bias [30, 32]. As a result, researchers rely on complementary strategies such as sensitivity analyses, negative controls, and triangulation across data sources or modeling approaches [13, 24].

Systematic Biases

Systematic biases in observational data come from three primary sources: confounding, selection bias, and measurement bias [24]. **Confounding**—widely recognized as the primary limitation of observational studies [28]—arises when common causes of treatment and outcome are not adequately controlled. **Selection bias** may result from non-random sampling or conditioning on colliders, while **measurement bias** reflects inaccuracies or incompleteness in observed variables.

From a causal graphical perspective, these biases correspond to open backdoor paths between treatment and outcome[24]. They affect both average treatment effects and heterogeneous effects, making conditional average treatment effect (CATE) estimation particularly sensitive to model misspecification and limited covariate overlap.

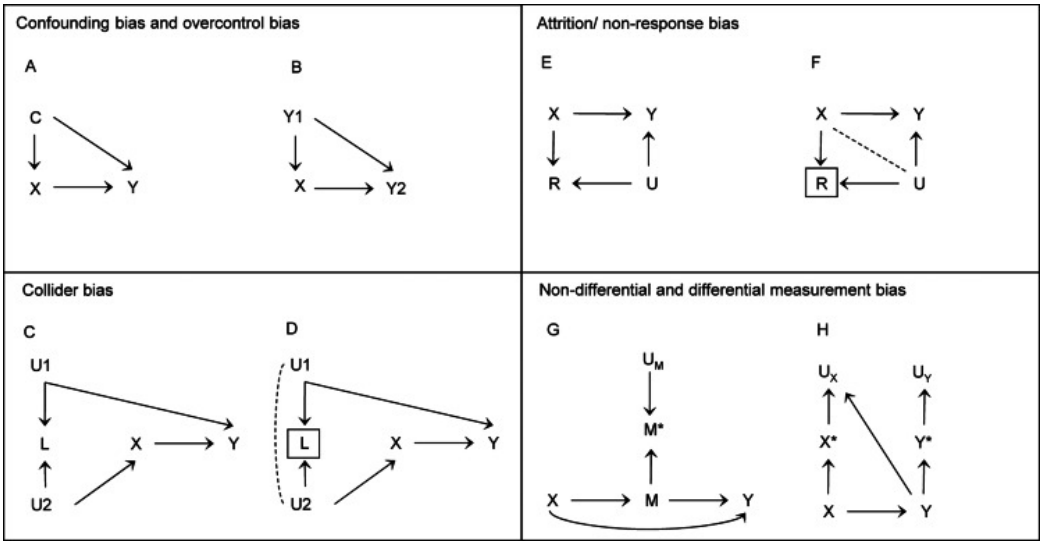

Figure 2.2: **Systematic biases in observational data illustrated as directed acyclic graphs (DAGs). (A–B) Confounding and overcontrol bias.** In **(A)**, an unmeasured or uncontrolled common cause C affects both X and Y, opening a backdoor path and inducing spurious association. In **(B)**, Y_1 is a mediator on the causal path from X to Y_2; conditioning on Y_1 constitutes overcontrol bias, blocking part of the true causal effect. **(C–D) Collider bias.** In **(C)**, two independent causes U_1 and U_2 both influence L, which in turn affects X; L acts as a collider, and conditioning on it opens a spurious path between U_1 and U_2, inducing bias. In **(D)**, L (boxed) is conditioned upon, activating the dashed path from U_1 to U_2 and introducing collider-stratification bias. **(E–F) Attrition and non-response bias.** In **(E)**, an unmeasured variable U jointly influences both Y and the retention indicator R, causing non-random missingness that distorts the observed $X \to Y$ relationship. In **(F)**, R (boxed, indicating conditioning) is a collider of X and U; restricting analysis to retained units ($R = 1$) opens a spurious association between X and Y via U (dashed arrow). **(G–H) Non-differential and differential measurement bias.** In **(G)**, the mediator M is measured with error as M^*, driven by an independent noise source U_M; this non-differential mismeasurement attenuates estimates of the $X \to M \to Y$ pathway. In **(H)**, both the treatment X and outcome Y are mismeasured as X^* and Y^* via independent noise sources U_X and U_Y respectively; when measurement error is differential (i.e. correlated with other variables), bias in the estimated $X \to Y$ effect can be amplified or reversed.

Confounding and Methods for Adjustment

As the main limitation of observational data [28], **confounding** becomes our center of interest for this project (which is also why we are artificially introducing it in our synthetic DGP to replicate observational data setting). The article by Hammerton [24] provides a systematic overview of statistical approaches that address confounding bias, which can be broadly grouped as follows:

- **Multivariable Regression**: Potential confounders are included in the regression model for the effect of the exposure on the outcome [25]
- **Propensity Estimation**: Propensity scores are used to control for time-invariant confounding, calculated by estimating the probability that an individual is exposed, given the values of their observed baseline confounders; can be extended to address time-varying confounding via marginal structural models [8]
- **Fixed Effects Regression**: This approach uses repeated measures of an exposure and an outcome to account for the possibility of an association between the exposure and the unexplained variability in the outcome (representing unmeasured confounding); can adjust for all time-invariant confounders, including unobserved confounders, and can incorporate observed time-varying confounders [16]

Modern frameworks for CATE estimation build directly on these adjustment-based approaches by combining outcome modeling and propensity estimation to recover heterogeneous treatment effects at the individual level. In particular, many recent methods reformulate causal estimation as a supervised learning problem, enabling the use of flexible machine learning models. These approaches are commonly referred to as *meta-learners*, which provide a unifying framework for leveraging standard predictive models for CATE estimation. The following section introduces the main classes of meta-learners and their underlying principles.

2.3 Modern Methods for CATE Estimation

2.3.1 Meta-Learning Framework

Modern methods for estimating CATE from observational data can be united by the general meta-learner framework. Meta-learners provide a unifying framework that casts CATE estimation as a supervised learning problem. Rather than relying on a specific model class, they are nonparametric in nature and can be combined with arbitrary machine learning methods [2, 37].

The meta-learning framework for CATE estimation was formalized by Künzel et al. [37], who introduced three fundamental approaches for the binary treatment setting: the S-learner, T-learner, and X-learner. These methods share a common structure: they first estimate *nuisance components*, namely the outcome regression functions and, in some cases, the propensity score, and then combine these estimates to construct targets for learning the CATE function.

Most meta-learners follow this two-stage procedure. In the first stage, nuisance functions are estimated using flexible machine learning models. In the second stage, these estimates are used to construct either potential outcomes or pseudo-outcomes, which serve as targets for supervised learning of the treatment effect.

Within this framework, indirect methods such as the S-learner and T-learner can be viewed as plug-in estimators that compute treatment effects as differences between predicted outcomes. More advanced learners refine this idea by directly targeting the CATE. The X-learner improves efficiency by imputing treatment effects and reweighting them using propensity scores [37].

Subsequent work has extended this framework to address key challenges such as model misspecification and statistical efficiency. The DR-learner, introduced by Kennedy [34], constructs doubly robust pseudo-outcomes that remain consistent if either the propensity score or the outcome model is

correctly specified. The R-learner [46] estimates the CATE by minimizing an orthogonalized loss function, which reduces sensitivity to nuisance estimation errors.

Additional variants include the PW-learner (propensity weighting, also known as the M-learner) and the RA-learner (regression adjustment), which can be interpreted as refinements of earlier approaches [11]. Notably, under suitable conditions, the RA-, PW-, and DR-learners can achieve oracle convergence rates, matching the performance of an estimator with access to true nuisance functions.

From a theoretical perspective, meta-learners differ primarily in their statistical efficiency and robustness to model misspecification. In particular, doubly robust learners exhibit favorable convergence properties and often outperform simpler plug-in estimators in terms of mean squared error. These advantages have been demonstrated both theoretically and empirically in [12].

2.3.2 Taxonomy of Meta-Learning Methods

Meta-Learning CATE estimation methods can be broken down into *indirect* or *direct* approaches. Indirect methods estimate potential outcome functions and compute CATE as their difference, whereas direct methods construct pseudo-outcomes that explicitly target the treatment effect [12].

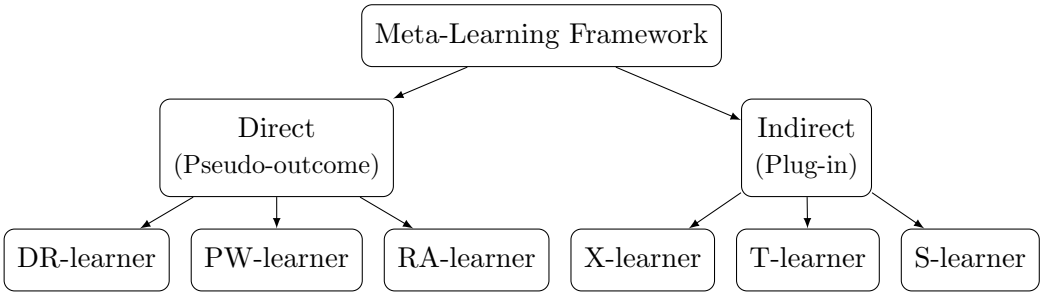

Figure 2.3: Taxonomy of Meta-Learners

Indirect Methods

Indirect approaches, often referred to as *plug-in estimators*, rely on modeling the outcome under treatment and control separately. Classical examples include the T-learner and S-learner, as well as more advanced meta-learning extensions [37].

The **T-learner** fits two separate models—one for the treated group and one for the control group—and estimates CATE as the difference between their predictions. While conceptually simple, this approach fails to fully exploit shared structure between the two groups, which can result in high variance, especially in small or imbalanced datasets [37, 64].

The **S-learner** instead fits a single model by including the treatment indicator as an additional covariate. Potential outcomes are obtained by evaluating the model under different treatment values. Although more data-efficient, this approach can suffer from bias if the model prioritizes predictive covariates over the treatment indicator, effectively underestimating treatment effects [64].

To address these limitations, more advanced meta-learning strategies have been proposed. The **X-learner** [37] augments outcome modeling with imputed treatment effects. It first estimates potential outcomes, then constructs pseudo-treatment effects for each group, and finally learns a model for these effects. By incorporating propensity-based weighting, the X-learner is particularly effective in settings with imbalanced treatment assignment.

A related line of work formulates CATE estimation as a representation learning or domain adaptation problem. **Domain adaptation learners** aim to learn balanced representations of covariates such that the distributions of treated and control groups become similar [54]. This reduces bias due to covariate shift and improves generalization of treatment effect estimates.

Despite their flexibility, indirect methods do not explicitly regularize the CATE function itself, as treatment effects are obtained implicitly as differences between outcome models.

Direct Methods

Direct methods estimate CATE by constructing pseudo-outcomes that explicitly target the treatment effect. Common approaches include regression-adjusted (RA), propensity-weighted (PW), and doubly robust (DR) estimators [12].

Propensity-weighted methods rely on inverse probability weighting to correct for selection bias. When the propensity score is correctly specified, the resulting pseudo-outcome is an unbiased estimator of the true CATE.

Doubly robust (DR) methods, such as the augmented inverse probability weighting (AIPW) estimator, combine outcome regression and propensity score modeling[10]. These estimators are particularly appealing because they remain consistent if *either* the propensity model or the outcome model is correctly specified, providing robustness against model misspecification.

2.3.3 Neural Network Implementations

As universal function approximators, neural networks naturally integrate into the CATE estimation framework by modeling the required nuisance components, such as outcome regressions $\mu_w(x)$ and propensity scores $\pi(x)$. In the simplest setting, each nuisance function is estimated independently using a separate neural network. This approach, often referred to as a *TNet*, directly mirrors classical meta-learners such as the T-learner, and is asymptotically flexible as it allows for arbitrarily different functional forms across tasks [12].

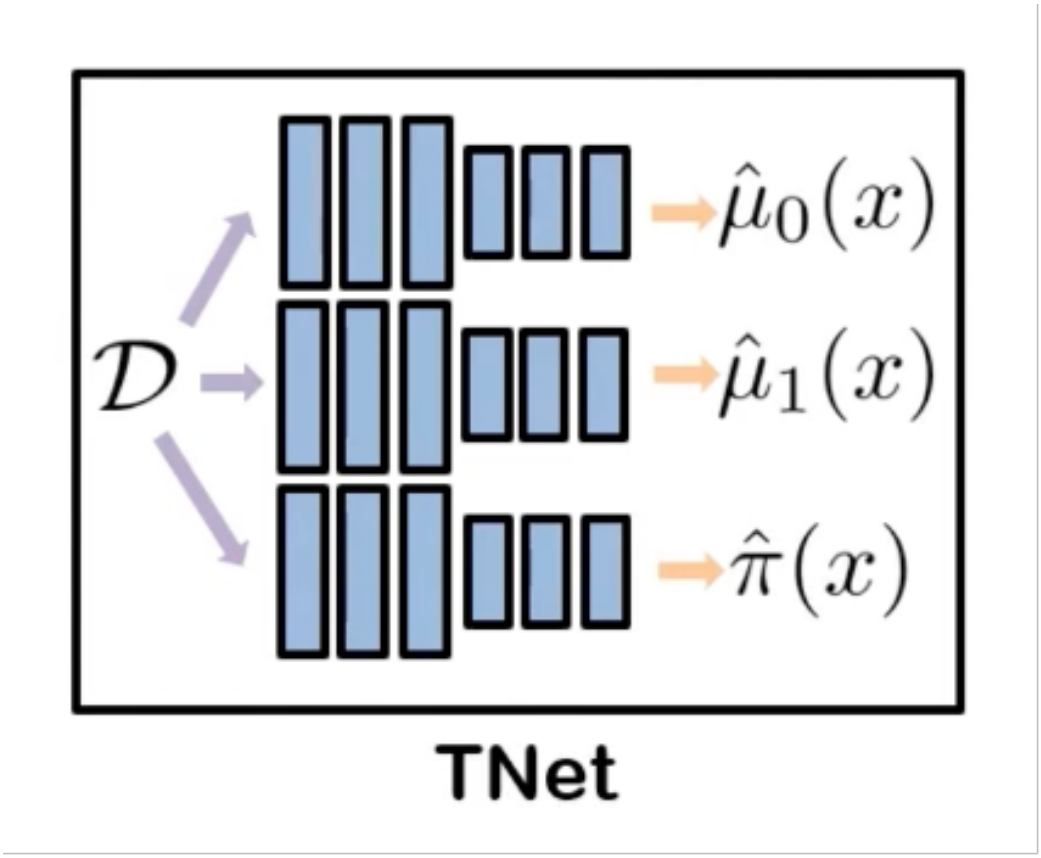

Figure 2.4: TNet architecture: independent neural networks for nuisance estimation. Separate task-specific heads are used to predict the potential outcomes $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$, as well as the treatment assignment probability $\hat{\pi}(x)$

However, independent estimation can be statistically inefficient in finite samples, as it fails to exploit potential similarities between nuisance functions. In many applications, the potential outcomes $\mu_0(x)$ and $\mu_1(x)$ are supported on similar covariate distributions and differ only through relatively simple treatment effects. In such cases, sharing information across tasks can substantially improve estimation accuracy.

To address this limitation, recent work proposes a class of architectures known as *shared representation networks* (SNets), which learn a common

feature representation $\Phi(x)$ across all tasks, followed by task-specific prediction heads. This approach enables the model to capture global structure in the data while retaining flexibility for treatment-specific effects, leading to improved statistical efficiency and generalization performance [12].

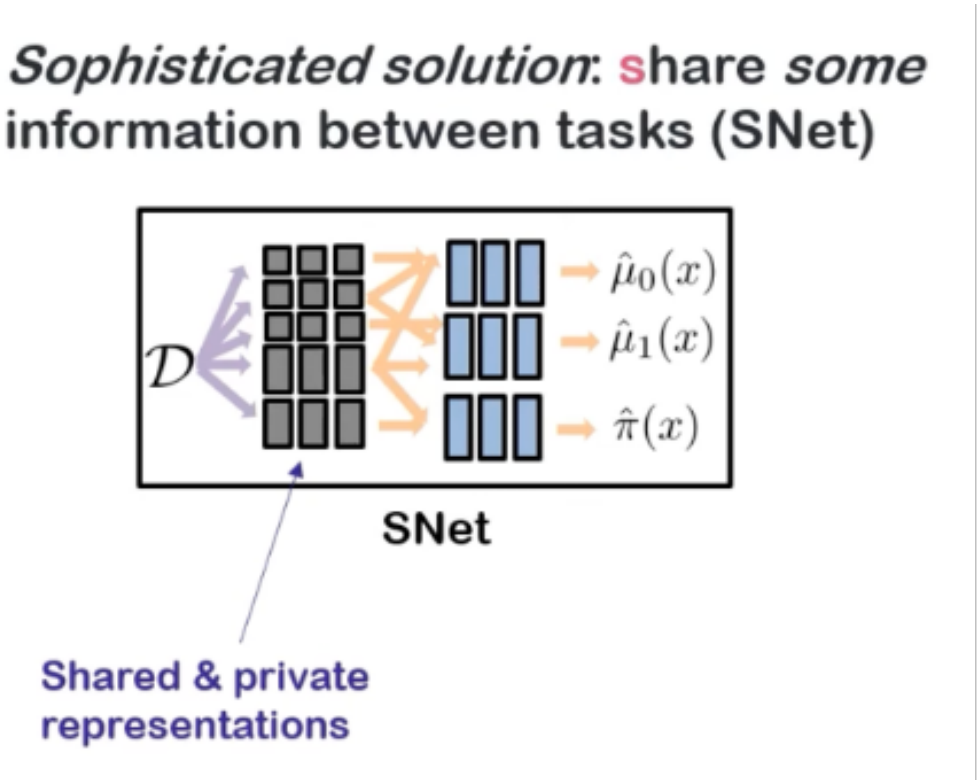

Figure 2.5: **SNet architecture for CATE estimation.** The model uses a combination of shared and private representations to capture both commonalities and task-specific features from the input data D. The shared representation (gray blocks) encodes information useful across multiple tasks, while the private representation captures task-specific nuances. These representations are fed into separate task-specific heads to predict the potential outcomes $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$, as well as the treatment assignment probability $\hat{\pi}(x)$. This design allows SNet to leverage information sharing between tasks while maintaining the flexibility to model individual task characteristics.

Several prominent architectures fall within this SNet framework. **TAR-**

Net [54] learns a shared representation with separate outcome heads for each treatment group. **DragonNet** [56] extends this idea by jointly modeling both outcome regressions and the propensity score, encouraging representations that balance treated and control groups. Finally, **DR-CFR** [27] incorporates doubly robust objectives and multiple representations to jointly capture outcome and propensity components, further improving robustness to model misspecification.

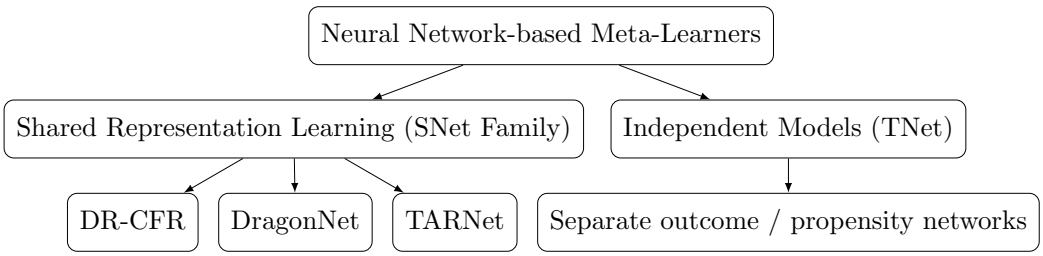

Figure 2.6: NN-based meta-learners. TNet corresponds to independent nuisance estimation, while SNet architectures share representations across tasks to improve efficiency.

Overall, neural network-based CATE estimation is best understood through the lens of how nuisance functions are parameterized and whether information is shared across tasks. While TNet architectures provide maximum flexibility, SNet architectures introduce an inductive bias that can significantly improve performance in finite samples. Combining meta-learning strategies (indirect or direct) with shared representation learning often yields strong empirical performance, particularly in high-dimensional settings [12].

2.4 Research Gap

Despite the growing body of work in EGL across various domains, **limited effort has been done to leverage EGL methods in causality**. To the best of our knowledge, [41] and [14] are the only examples in this direction.

EGL has been predominantly studied in areas such as computer vision (CV), natural language processing (NLP), and Visual Question Answering (VQA), showing benefits in terms of model explainability and generalizability to unseen data [18]. However, EGL applications in causal inference have been overlooked.

In parallel, recent research on CATE estimation has focused on improving the robustness of causal effect estimation under the limitations of observational data. Key directions include: (i) reducing bias under weaker ignorability assumptions through techniques like proximal causal learning [58, 62], (ii) handling limited covariate overlap and confounding via representation learning and balancing approaches [35], (iii) increasing efficiency and robustness using double/debiased machine learning and targeted maximum likelihood estimation [10], (iv) providing valid uncertainty quantification through conformal inference and honest estimation [42], and (v) improving policy-relevant decision making by integrating CATE estimates with treatment assignment optimization [33].

However, despite these advances, **CATE research has largely focused on statistical and algorithmic improvements rather than incorporating explainability or guidance from domain knowledge** in the estimation process. In particular, there is a lack of methods that utilize explanations to guide the estimation of nuisance components (e.g., outcome and propensity models), which are critical for accurate CATE estimation.

Consequently, there is a lack of controlled empirical studies that isolate the impact of explanation signals on causal estimation performance. In particular, it remains unclear under which conditions (e.g., strength of confounding, type of explanation, integration mechanism) EGL provides measurable benefits for CATE estimation. Addressing this gap requires a unified framework that combines causal inference with EGL, together with systematic experimentation.

This thesis aims to bridge these gaps by formalizing EGL within the

CATE estimation setting, proposing a mechanism for integrating explanation signals into causal learning objectives, and empirically evaluating their effectiveness under controlled confounding scenarios.

Chapter 3

Research Objectives

3.1 Problem Formulation

3.1.1 Observations with Explanations

Consider $\mathcal{X} \subset R^{d_x}$, $\mathcal{T} \subset R^{d_t}$, $\mathcal{Y} \subset R^{d_y}$, $\mathcal{E}_\mathcal{Y} \subset R^{d_{e_y}}$, $\mathcal{E}_\mathcal{T} \subset R^{d_{e_t}}$ as the spaces of covariates, treatments, outcomes, outcome explanations, and treatment explanations, respectively. Let $\mathbf{X} \in \mathcal{X}$, $\mathbf{T} \in \mathcal{T}$, $\mathbf{Y} \in \mathcal{Y}$, $\mathbf{E_Y} \in \mathcal{E}_\mathcal{Y}$, $\mathbf{E_T} \in \mathcal{E}_\mathcal{T}$ be the corresponding feature vectors. We assume the directed acyclic graph shown in Figure 3.1, together with $\mathbf{E_T} = f_{E_T}(\mathbf{X}, f_T)$ and $\mathbf{E_Y} = f_{E_Y}(\mathbf{X}, \mathbf{T}, f_Y)$.

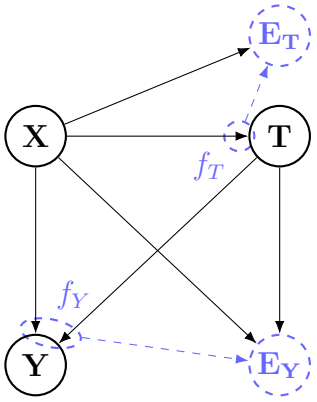

Figure 3.1: **Graphical representation of the assumed directed acyclic graph.** Node X represents covariates of a subject, node T stands for treatment allocation (treated or control), and Y represents the observed outcome. f_Y is the outcome ground truth function that depends on covariates and treatment, and f_T is the function of propensity assignment. E_Y depicts the explanations of outcomes, while E_T stands for explanations of propensity assignment.

Structural equation model corresponding to the DAG in Figure 3.1:

$$\begin{cases} \mathbf{X} \sim P_X \\ \mathbf{T} := f_T(\mathbf{X}) + \varepsilon_T \\ \mathbf{E_T} := f_{E_T}(\mathbf{X}, f_T) + \varepsilon_{E_T} \\ \mathbf{Y} := f_Y(\mathbf{X}, \mathbf{T}) + \varepsilon_Y \\ \mathbf{E_Y} := f_{E_Y}(\mathbf{X}, \mathbf{T}, f_Y) + \varepsilon_{E_Y} \end{cases}$$

3.1.2 CATE Estimation

Potential Outcomes Framework

We operate under the standard setup of the potential outcomes (PO) framework [28], where each subject, with covariates $\mathbf{X} \sim P_X$, has two potential outcomes: $Y(1)$ under treatment and $Y(0)$ under control, of which only one is observed. We consider the binary treatment setting $T \in \{0, 1\}$.

The goal is to estimate the Conditional Average Treatment Effect (CATE), defined as the expected difference between an individual's potential outcomes, conditional on covariates:

$$\tau(x) = E[Y(1) - Y(0) \mid X = x] = \mu_1(x) - \mu_0(x) \tag{3.1}$$

where $\mu_\omega(x) = E[Y(\omega) \mid X = x]$ denotes the expected potential outcome.

Identifying Causal Assumptions

Similar to experimental approaches in recent CATE literature [11], we rely on strong ignorability conditions [48] for convenience, while acknowledging potential limitations in practical applications. When it comes to assumptions on the potential outcomes, we assume that POs share a common baseline, with the treated outcome modeled as an additive combination of the control outcome and a heterogeneous treatment effect,

$$Y = \mu_1(X) \cdot T + \mu_0(X) \cdot (1 - T) + \epsilon \tag{3.2}$$

, though more general transformations are possible [12, 37].

Integrating EGL Component into the Objective Function

We focus on improving the regression surface component of the meta-learners via EGL. The outcome regression component is indicated with a) in the diagram below for the X-Learner type [37]:

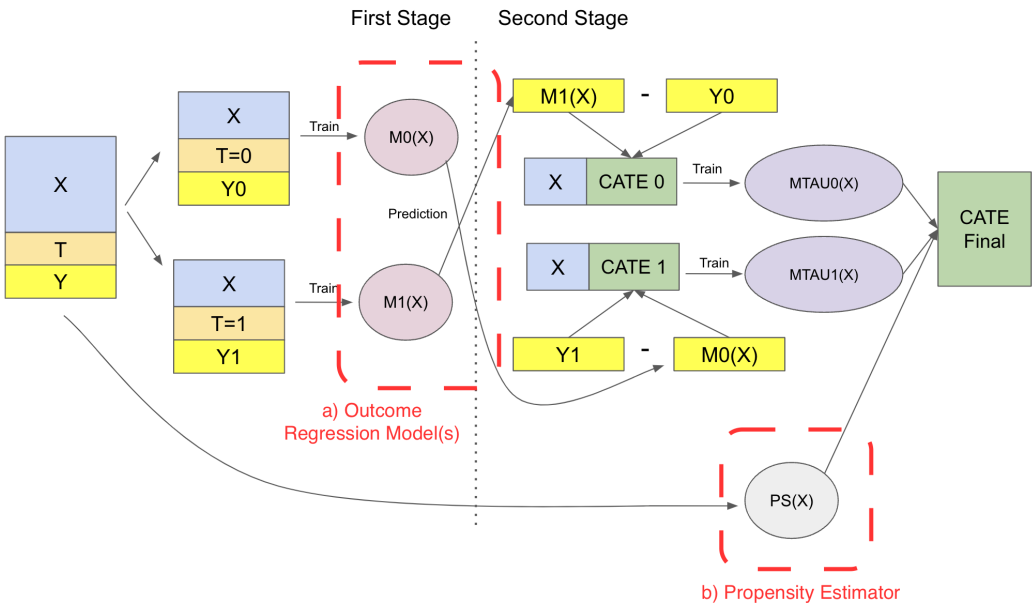

Figure 3.2: Components of an X-Learner. Both a) Outcome Regression and b) Propensity Estimator components can benefit from EGL injection. In this thesis, however, we target a) only.

To incorporate additional structural knowledge, we assume access to explanations E describing how input features should influence outcome estimation. These explanations may represent, for example, feature attribution scores, counterfactual samples, or binary relevance masks indicating which variables are expected to affect treatment assignment or outcomes. In our experiments, we focus on **counterfactual explanations** and rely on DiCE [45] as the explanation generator. We denote the learned outcome model by $\hat{\mu}(X, T)$, which estimates the expected outcome given covariates and treatment. The explainer can then be represented as a function g that takes the learned model and an input (X, T) and produces a counterfactual explanation:

$$g(\hat{\mu}, X, T) = \hat{E}_{\text{cf}} = (X_{\text{cf}}, T_{\text{cf}}, Y_{\text{cf}}) \tag{3.3}$$

where $(X_{\text{cf}}, T_{\text{cf}})$ is a perturbed input and Y_{cf} is the corresponding counter-

factual outcome generated by the oracle model.

We incorporate these expert explanations through a data augmentation approach. The resulting training objective combines the standard predictive loss on observational data with an additional loss defined on counterfactual samples:

$$\min_{\hat{\mu}} \ \underbrace{\mathcal{L}_{\text{pred}}\big(\hat{\mu}(X,T),Y\big)}_{\text{factual loss}} + \lambda \ \underbrace{\mathcal{L}_{\text{exp}}\big(\hat{\mu}(X_{\text{cf}},T_{\text{cf}}),Y_{\text{cf}}\big)}_{\text{counterfactual loss}}. \tag{3.4}$$

Here:

- $\mathcal{L}_{\text{pred}}$ is the predictive loss (e.g., MSE) computed on observed (factual) data,
- $\mathcal{L}_{\text{exp}}$ is the predictive loss evaluated on counterfactual samples,
- λ controls the relative contribution of the explanation-guided (counterfactual) loss.

By incorporating explanation-based supervision into the training process, EGL provides a mechanism for injecting domain knowledge into the learning process, encouraging the model to rely on causally relevant covariates.

Performance Metric Following prior work on heterogeneous treatment effect estimation [54], we evaluate models using the **mean absolute error (MAE) in CATE**.

$$\epsilon_{\text{CATE}} = \frac{1}{n}\sum_{i=1}^{n} \left|(\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)) - (\mu_1(x_i) - \mu_0(x_i))\right| \tag{3.5}$$

Here, $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ denote the *estimated potential outcome functions* learned by the model, while $\mu_1(x)$ and $\mu_0(x)$ correspond to the *ground-truth outcome functions* defined by the data-generating process.

The quantity $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$ represents the predicted treatment effect, and $\tau(x) = \mu_1(x) - \mu_0(x)$ is the true treatment effect. The metric ϵ_{CATE}

therefore measures the average absolute deviation between the estimated and true CATE across all samples.

3.2 Research Questions

The goal of this thesis is to empirically explore how explanation signals influence the accuracy of causal effect estimation. We focus on the following questions:

- How can human- or model-generated explanations be integrated into causal effect estimation frameworks?
- How effective are counterfactual explanations for EGL in CATE estimation?
- Which EGL technique (type of guidance) is most feasible with Meta-Learners?

Our experiments provide novel insights into when and how explanation signals can be used as a principled auxiliary source of information to improve causal effect estimation. This work represents a step toward integrating explanations into causal machine learning, with potential applications in healthcare, economics and policy analysis, marketing and customer analytics, and other causal domains.

Chapter 4

Experimental Setup

4.1 Data

Following common practices in causal inference literature [12, 37], we adopt a *fully controlled synthetic setup* to evaluate the effectiveness of EGL for CATE estimation. This setting allows us to control the level of confounding, generate randomized (RCT) test data, access ground-truth treatment effects, and precisely quantify estimation bias.

Our synthetic data generating process follows [37] and is defined within the potential outcomes framework:

$$Y = \mu_1(X) \cdot T + \mu_0(X) \cdot (1 - T) + \epsilon, \quad T \sim \text{Bernoulli}(e(X)), \quad \epsilon \sim \mathcal{N}(0, 1) \tag{4.1}$$

where $e(X) = P(T = 1 \mid X)$ is the propensity score.

Covariates are initially sampled from a multivariate normal distribution:

$$X \sim \mathcal{N}(0, I_d), \quad d = 5 \tag{4.2}$$

To facilitate a detailed analysis of treatment effect heterogeneity, we override

the second covariate (X_2) with a deterministic grid:

$$X_2 \in [-3, 3]$$

Specifically, X_2 is evenly spaced across this interval, while all remaining covariates retain their stochastic Gaussian structure.

4.1.1 Outcome Model Specification

The baseline outcome function is linear in the covariates:

$$\mu_0(X) = X^\top \beta, \quad \beta \sim \text{Unif}([-3, 3]^d) \tag{4.3}$$

The treated outcome is defined as:

$$\mu_1(X) = \mu_0(X) + 8 \cdot I(X_1 > 0.1) \tag{4.4}$$

This induces a heterogeneous treatment effect:

$$\tau(X) = \mu_1(X) - \mu_0(X) = 8 \cdot I(X_1 > 0.1) \tag{4.5}$$

Thus, the observed outcome can equivalently be written as:

$$Y = X^\top \beta + 8 \cdot I(X_1 > 0.1) \cdot T + \epsilon \tag{4.6}$$

4.1.2 Training Set

Our training dataset is explicitly generated with confounding to replicate the common challenge of observational data. The confounding mechanism is defined through a deterministic propensity function based on covariate X_2:

$$e(x) = P(T = 1 \mid X) = \begin{cases} 0.95 & \text{if } -2 < x_2 < 2 \\ 0.05 & \text{otherwise} \end{cases} \tag{4.7}$$

This creates strong selection bias: units with moderate values of X_2 are almost always treated (95% probability), while units with extreme X_2 values are rarely treated (5% probability). Importantly, the confounding variable X_2 is *different* from the effect modifier X_1, creating a challenging estimation problem.

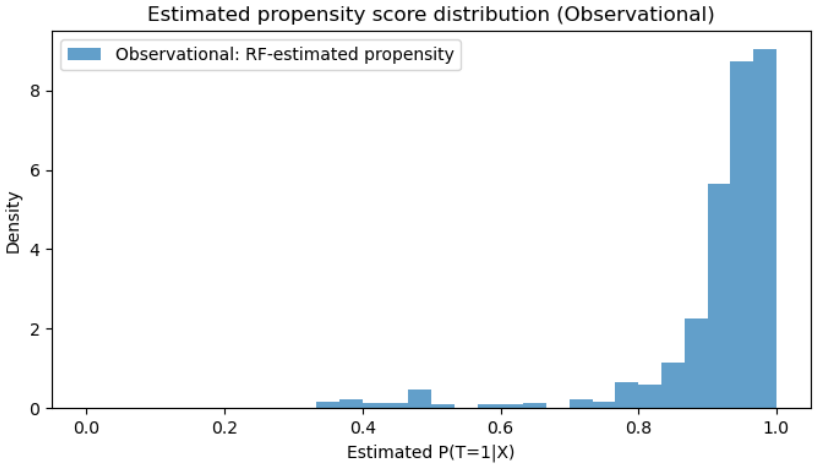

Figure 4.1: Propensity score distribution of observational training data, showing strong selection bias with most units having propensity near 0.95.

4.1.3 Test Set

The test set is generated under a RCT setting, where treatment assignment is independent of covariates:

$$e_{\text{RCT}}(X) = 0.5 \quad \forall X, \quad T \sim \text{Bernoulli}(0.5) \tag{4.8}$$

This removes confounding and enables unbiased evaluation of treatment effect estimates.

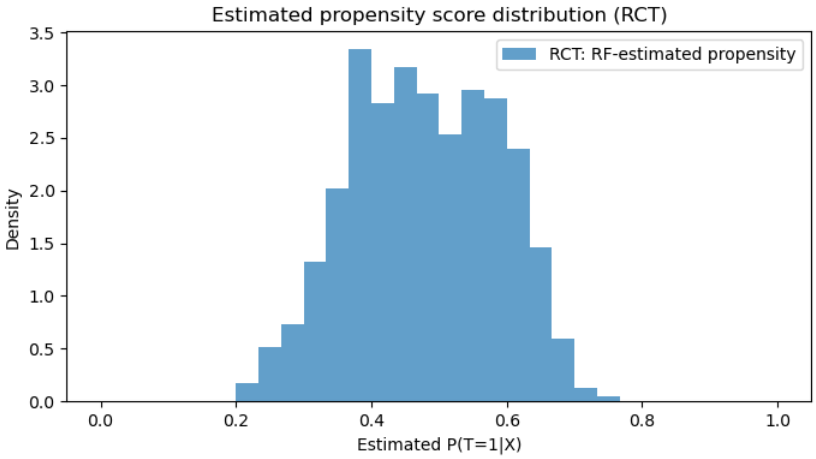

Figure 4.2: Propensity score distribution of RCT test data, showing uniform distribution around 0.5.

4.1.4 Ground Truth Explanations: Counterfactuals

We generate "expert" explanations as counterfactuals using the DiCE framework [45]. These counterfactuals are constructed using the oracle outcome function, allowing us to produce unbiased signals for data augmentation.

Oracle Model To generate counterfactual outcomes, we wrap the ground-truth data-generating function $f^*(X, T)$ as a predictive model compatible with DiCE. This oracle model takes as input both covariates and treatment:

$$\hat{Y} = f^*(X, T) \tag{4.9}$$

This ensures that all generated counterfactual outcomes are **consistent with the true underlying data-generating process.**

Counterfactual Generation Procedure Counterfactuals are generated using DiCE with the following configuration:

- **Features allowed to vary:** all features

- **Constraint:** treatment is restricted to the flipped value
- **Number of counterfactuals:** 1 per instance
- **Outcomes:** The outcome range is preserved according to the range in the observational training set

DiCE generates counterfactuals by solving an optimization problem that searches for small perturbations of the input features which lead to a desired change in the model output. Specifically, it balances three objectives: (i) *validity*, ensuring the counterfactual achieves the desired outcome; (ii) *proximity*, keeping the counterfactual close to the original instance; and (iii) *feasibility*, enforcing constraints on which features can change and how.

In our setting, since the outcome variable is continuous, we also treat Y_{CF} as continuous and allow covariates to adjust to produce a valid counterfactual.

This produces a complete counterfactual dataset: $(\mathrm{X}_{\mathrm{CF}}, T_{\mathrm{CF}}, Y_{\mathrm{CF}})$

These counterfactual samples serve as *expert explanations of outcomes*, as they explicitly describe how the outcome would change under an alternative treatment assignment. Unlike observational data, these explanations exhibit reduced confounding, since they are generated from the oracle model.

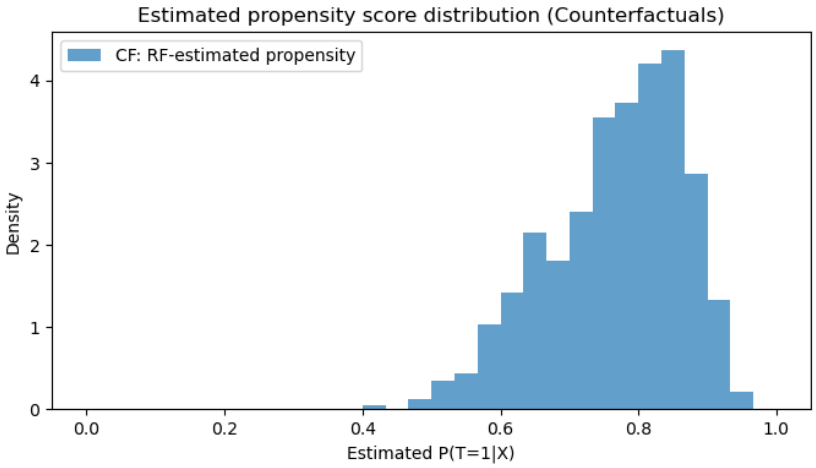

Figure 4.3: Propensity score distribution of counterfactual explanations. The distribution is shifted closer to that of RCT compared to the observational one, showcasing the potential of debiasing causal models with this new signal.

We then integrate this set of "expert" counterfactuals into the training of outcome regression nuisance models that most meta-learners use.

4.2 Methods

4.2.1 Selected Meta-Learners

The following meta-learners were selected for our experiments:

- **S-Learner**: A single model $\hat{\mu}(X, T)$ that predicts outcomes given covariates and treatment. CATE is estimated as:

$$\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0) \tag{4.10}$$

- **X-Learner**: A two-stage learner that first estimates outcome models, then fits separate CATE models on imputed treatment effects for

treated and control groups, combining them via propensity weighting:

$$\hat{\tau}(x) = e(x) \cdot \hat{\tau}_0(x) + (1 - e(x)) \cdot \hat{\tau}_1(x) \tag{4.11}$$

- **DR-Learner**: A doubly robust learner that computes pseudo-outcomes using the AIPW formula:

$$\Gamma_i = (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{e(X_i)} - \frac{(1 - T_i)(Y_i - \hat{\mu}_0(X_i))}{1 - e(X_i)} \tag{4.12}$$

  then fits a final CATE model on these pseudo-outcomes.

- **DA-Learner**: A domain adaptation learner that uses importance weighting via propensity scores to reweight samples, focusing on the overlap region using weights $w(x) = \min(e(x), 1 - e(x))$.

As mentioned in previous chapters, all the above meta-learners rely on regression surfaces for outcome prediction. Therefore, we train a single outcome model with EGL loss and use it as a plug-in component for various architectures. We then assess how the EGL contributes to the final result of CATE estimation for different types of meta-learners.

The base outcome regression model is implemented with a neural network with the following hyperparameters chosen according to standard deep learning practices [20]. Since we only care about the difference in performance of our outcome models (with and without EGL), tuning these hyperparameters to achieve the highest starting accuracy was not needed.

Parameter	Value
Hidden dimensions	64
Activation function	ReLU
Number of layers	2
Learning rate	10^{-3}
Batch size	64
Maximum epochs	200
Early stopping patience	150
Optimizer	Adam

Table 4.1: Neural network hyperparameters for the S-learner base model. Since we only care about the difference in performance of our outcome models (with and without EGL), tuning these hypermarameters to achieve the highest starting accuracy was not needed.

All experiments use the same random seed (123) for reproducibility, and validation is performed on RCT data to select the best model via early stopping. For the X-learner, DR-learner, and DA-learner architectures, we use an additional nuisance model – a propensity estimator. In our experiments, it is implemented via Random Forest regression model.

4.2.2 Implementation of EGL in Meta-Learners

We integrate the explanation signal using a data augmentation approach [18]. Expert explanations are incorporated into the training process via a separate counterfactual loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{factual}} + \lambda_{\text{CF}} \cdot \mathcal{L}_{\text{counterfactual}} \tag{4.13}$$

where:

- $\mathcal{L}_{\text{factual}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}(X_i, T_i))^2$ is the standard MSE loss on observed data,
- $\mathcal{L}_{\text{counterfactual}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i^{\text{CF}} - \hat{\mu}(X_i, T_i^{\text{CF}}))^2$ is the MSE on counterfactual examples,

- λ_{CF} controls the strength of explanation supervision. It is not bounded, but we keep it in the range [0, 100] when searching for the optimal value.

S-Learner with EGL The S-learner receives (X, T) as input and is supervised on both factual and counterfactual outcomes simultaneously. During each training iteration, the model processes batches from both the observational data and the counterfactual explanations, computing separate losses for each.

X-Learner, DR-Learner, and DA-Learner with EGL For the X-learner, DR-learner, and DA-learner, we apply EGL to the underlying S-learner that serves as the nuisance outcome model $\hat{\mu}(X, T)$. The improved outcome predictions from the EGL-augmented S-learner then propagate through the meta-learner pipeline:

- **X-Learner**: Better $\hat{\mu}_0, \hat{\mu}_1$ estimates lead to more accurate imputed treatment effects $D_1 = Y - \hat{\mu}_0(X)$ and $D_0 = \hat{\mu}_1(X) - Y$.
- **DR-Learner**: Improved outcome model reduces the residual terms $(Y - \hat{\mu}_1)$ and $(Y - \hat{\mu}_0)$ in the AIPW pseudo-outcome.
- **DA-Learner**: Better outcome predictions directly improve the plug-in CATE estimate $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$.

Chapter 5

Results and Discussion

The introduction of counterfactual data augmentation through EGL led to a **consistent reduction in mean absolute bias across all evaluated meta-learners**. However, the magnitude of this improvement varied substantially between architectures, revealing important differences in how explanation signals are propagated through the learning pipeline.

S-Learner

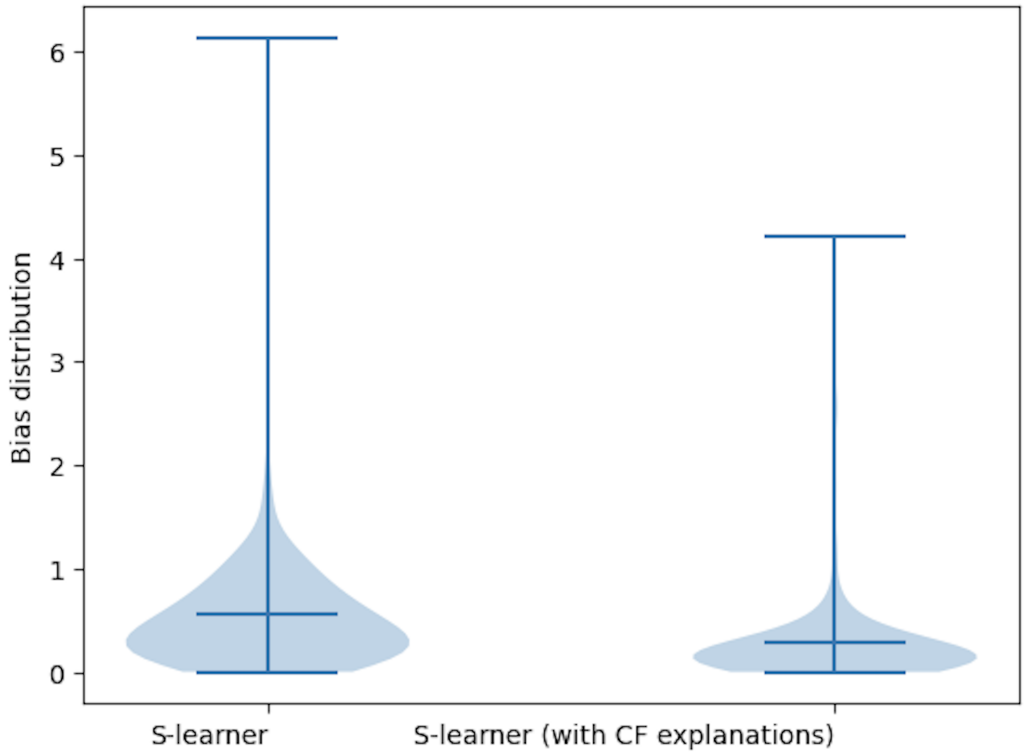

Figure 5.1: Distribution of absolute bias for the S-Learner. EGL reduces both the mean and variance of estimation error, indicating improved stability of the learned CATE. **Mean Absolute Bias Decrease of 48%**

X-Learner

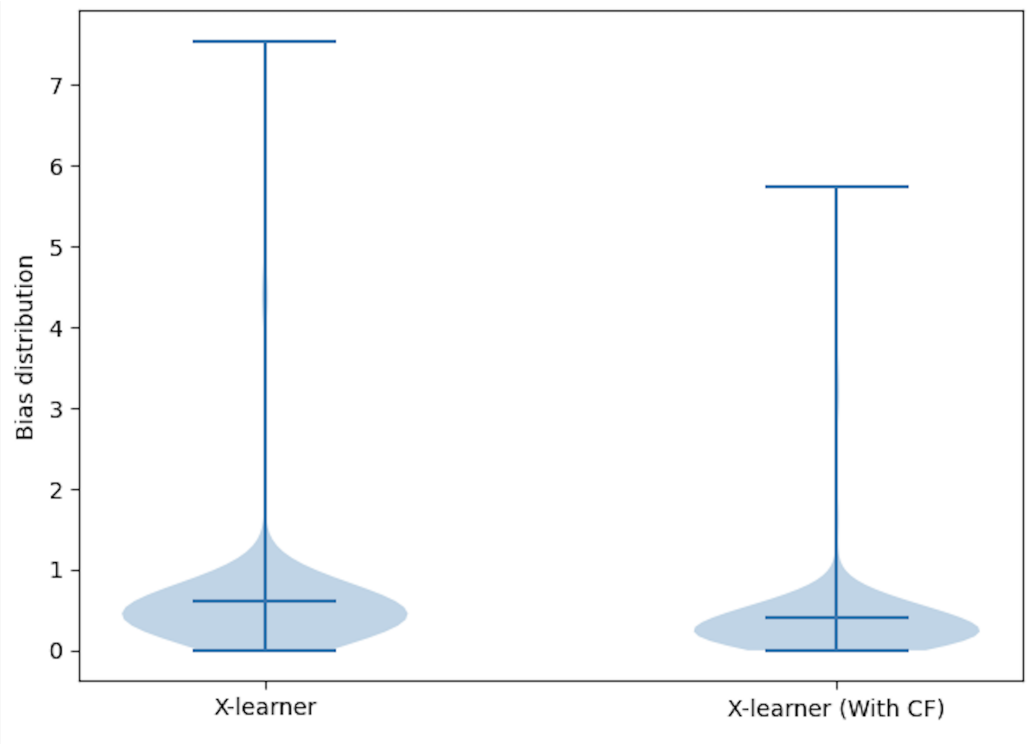

Figure 5.2: Distribution of absolute bias for the X-Learner. Improvements are visible, reflecting better estimation of imputed treatment effects through enhanced outcome models. **Mean Absolute Bias Decrease of 31%**

DR-Learner

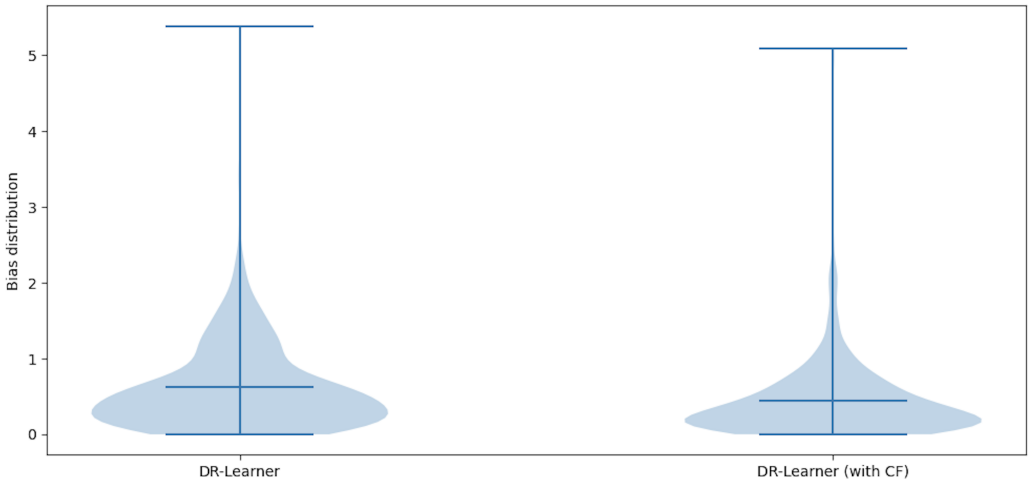

Figure 5.3: Distribution of absolute bias for the DR-Learner. While EGL reduces bias, the improvement is less pronounced compared to other architectures. **Mean Absolute Bias Decrease of 30%**

DA-Learner

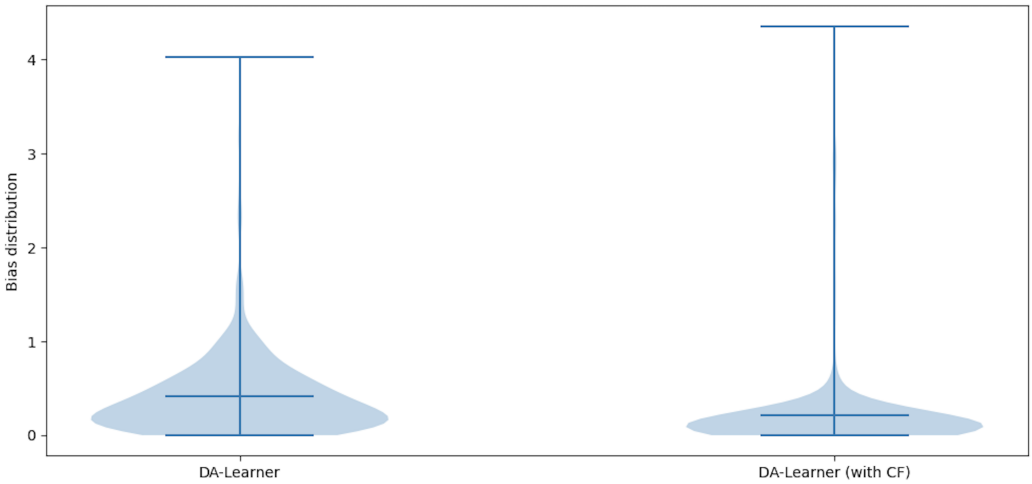

Figure 5.4: Distribution of absolute bias for the DA-Learner. The strongest improvement is observed, suggesting that this architecture is particularly well-suited to leveraging EGL signals in the context of our DGP. **Mean Absolute Bias Decrease of 49%**

Meta-Learner	Mean Absolute Bias Decrease
S-learner	48%
X-learner	31%
DR-learner	30%
DA-learner	49%

Table 5.1: Aggregated Results: Decrease in Mean Absolute Bias per Meta-Learner Variant

Most notably, the DA-Learner exhibited the largest improvement, achieving approximately a 49% reduction in mean bias (from 0.4161 to 0.2103). In contrast, the DR-Learner, while still benefiting from EGL, showed a smaller relative reduction of around 30% (from 0.6263 to 0.4395).

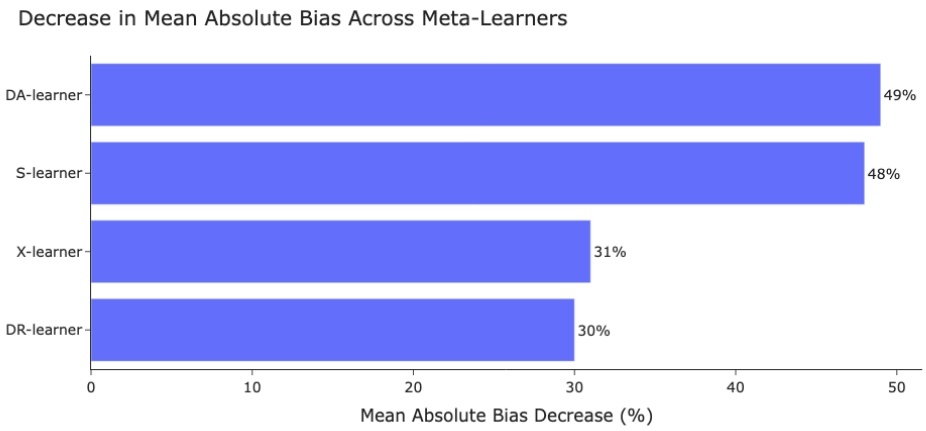

Figure 5.5: Decrease in Mean Absolute Bias across Meta-Learner Architectures

Tuning the Influence of EGL: $\lambda_{\mathbf{CF}}$ parameter

We explored how different values of λ_{CF}—the parameter controlling the strength of explanation supervision—affect the performance of our causal estimators. To find an optimal value, we followed a simple approach of re-running the experiments for each value of λ_{CF} in the set [0, 0.1, 0.5, 1, 2, 5, 10, 20, 50, 70, 100]. For example, when $\lambda_{\text{CF}} = 0$, the model is trained only on factual observational data (baseline). When $\lambda_{\text{CF}} = 5$, counterfactual explanations contribute significantly to the training signal.

The dynamic of different values of λ_{CF} is demonstrated in the figures below for S-learner and X-learner:

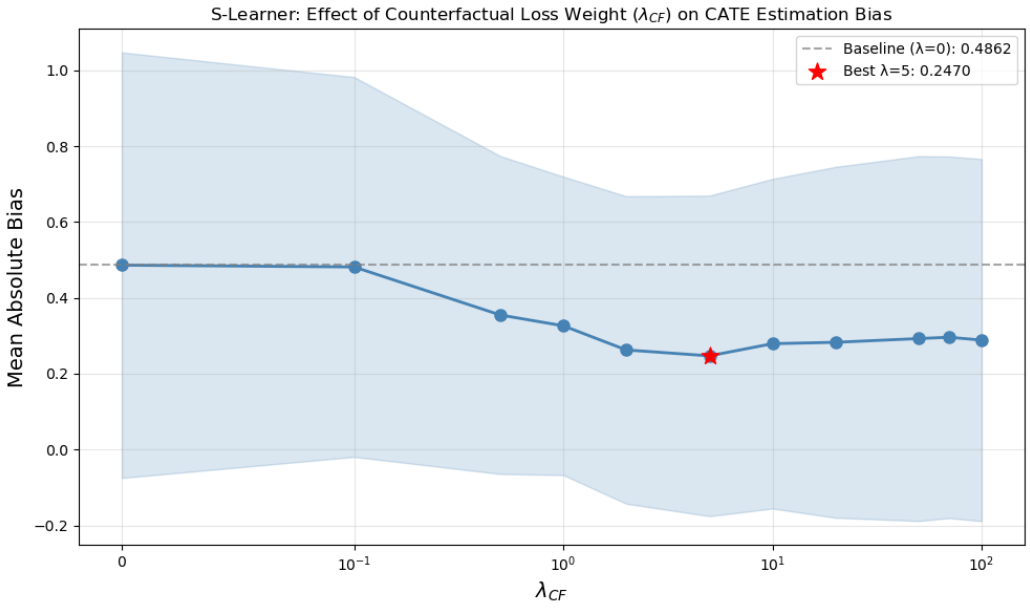

Figure 5.6: Effect of λ_{CF}
on Mean Absolute Bias in S-Learner

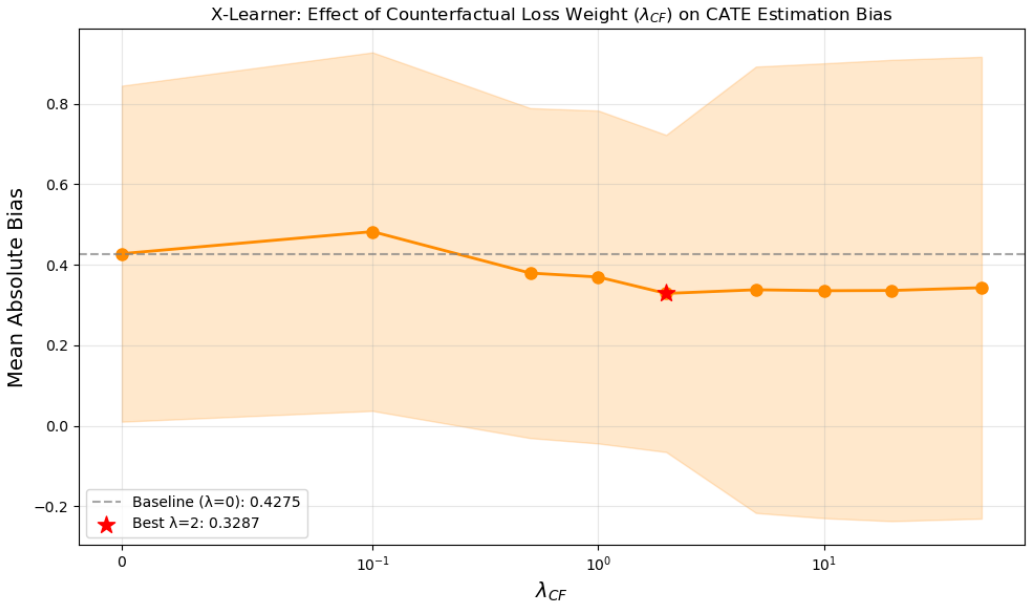

Figure 5.7: Effect of λ_{CF}
on Mean Absolute Bias in X-Learner

The effect of the counterfactual loss exhibits a non-monotonic relationship with bias reduction: performance improves as $_{CF}$ increases up to a moderate range, after which further increases lead to diminishing returns.

Chapter 6

Conclusions

6.1 Contributions

The main novelty and contribution of this thesis lie in bridging the gap between causal inference and EGL. To the best of our knowledge, EGL techniques have so far been primarily applied in domains such as computer vision, NLP, Visual Question Answering (VQA), graph-based learning, and purely associative predictive tasks [18] – but not in causal inference. Introducing EGL to this novel domain and systematically evaluating its effectiveness through varied experiments, therefore, leads to a substantial contribution. In a more detailed way, the contributions can be summarized as follows:

- **First Steps in Formalization of EGL in causal inference.** We extend the EGL paradigm to the setting of CATE estimation by explicitly incorporating explanations as auxiliary signals within the causal learning pipeline. We formalize a data-generating process that incorporates outcome explanations, present a corresponding graphical representation of the DAG, and provide a set of structural equations describing the underlying causal structure.

- **Integration of explanation signals into CATE estimation ob-**

jectives. We explicitly define an objective function with EGL loss with a data augmentation approach. Specifically, we augment standard predictive loss with an explanation-based loss defined on counterfactual samples, enabling the model to leverage additional structural information about the outcome-generating process. We demonstrate how EGL can be integrated into widely used meta-learners (S-, X-, DR-, and DA-learners) through the outcome regression component. By treating the EGL-enhanced model as a plug-in estimator, we systematically evaluate its impact across different causal estimation frameworks.

- **Empirical analysis of EGL effectiveness.** We provide empirical evidence on how explanation signals improve causal effect estimation by demonstrating a reduction in mean absolute bias. We demonstrate their role as an auxiliary source of information for mitigating bias in outcome models under confounding.

6.2 Future Research

In the current iteration of this work, we consider post-hoc explanations of observed outcomes as the primary "expert signal", which is integrated into outcome regression models within causal meta-learners. A natural extension is to incorporate **explanations of treatment assignment**, which we expect to be readily interpretable and accessible in practice. Leveraging such explanations would allow us to de-bias not only the outcome nuisance components, but also the **propensity models**. When both outcome- and treatment-related expert signals are available, we expect improvements in CATE estimation to be even more robust.

So far, our approach has focused on EGL-based data augmentation. In future work, we plan to investigate **alternative modes of EGL integration**, in particular through **supervision and regularization**. These approaches aim to align expert-provided explanations with model-derived interpretabil-

ity signals, thereby guiding the model toward more expert-consistent reasoning. Concretely, this requires selecting suitable explainers for generating model-side explanations and ensuring compatibility with the representation of human-provided signals. For instance, when expert knowledge is expressed in terms of feature importance, one could employ SHAP-based explanations [9]. Systematically evaluating the effectiveness of such off-the-shelf explainers remains an important direction for future work.

A key practical consideration in this context is whether the chosen **explainer is differentiable**, enabling its use within gradient-based optimization pipelines. Many popular explainers, such as LIME [47], are not inherently differentiable, which complicates their integration into EGL-based supervision schemes. This limitation motivates the exploration of surrogate approximations or gradient-based methods [52].

Beyond meta-learning approaches, we also aim to study how EGL techniques can be incorporated into **alternative CATE estimation frameworks**, particularly deep learning–based models such as DragonNet [55], CFRNet [31], and Deep Counterfactual Networks with Propensity Dropout (DCN-PD) [4]. These architectures rely on representation learning and implicit balancing mechanisms, making them a promising setting for integrating expert-guided signals at the representation or regularization level.

Finally, the current experimental setup is limited to fully synthetic data and does not account for **hidden confounding**, which restricts its realism. As part of future work, we plan to incorporate hidden confounding into the data-generating process and conduct sensitivity analyses, following established approaches in the causal inference literature [28, 49].

In addition, we aim to extend our evaluation to semi-synthetic datasets, which provide a more realistic benchmark while retaining partial ground truth. This will enable a more rigorous assessment of the robustness and practical applicability of EGL-enhanced estimators under conditions that more closely resemble real-world data.

Bibliography

[1] Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. L 2024/1689.

[2] Naoufal Acharki, Ramiro Lugo, Antoine Bertoncello, and Josselin Garnier. Comparison of meta-learners for estimating multi-valued treatment heterogeneous effects. *arXiv preprint arXiv:2303.XXXXX*, 2023.

[3] Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[4] Ahmed M. Alaa, Michael Weisz, and Mihaela van der Schaar. Deep counterfactual networks with propensity-dropout, 2017. URL `https://arxiv.org/abs/1706.05966`.

[5] Ahmed Aloui, Juncheng Dong, Ali Hasan, and Vahid Tarokh. Conditional average treatment effect estimation under hidden confounders, 2025. URL `https://arxiv.org/abs/2506.12304`.

[6] Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. doi: 10.1080/01621459.1997.10474074.

[7] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, 2024. doi: 10.1109/TNNLS.2022.3229161.

[8] Bethany C Bray, John J Dziak, Megan E Patrick, and Stephanie T Lanza. Inverse propensity score weighting with a latent class exposure: Estimating the causal effect of reported reasons for alcohol use on problem alcohol use 16 years later. *Prevention Science*, 20(3):394–406, 2019.

[9] Jianbo Chen et al. Shapnn: Shapley value-based neural networks for tabular data. *arXiv preprint arXiv:2206.XXXX*, 2022.

[10] Victor et al. Chernozhukov. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 2018.

[11] Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation, 2021. URL `https://arxiv.org/abs/2106.03765`.

[12] Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms, 2021. URL `https://arxiv.org/abs/2101.10943`.

[13] Issa J. Dahabreh and Kirsten Bibbins-Domingo. Causal inference about the effects of interventions from observational studies in medical journals. *JAMA*, 331(21):1845–1853, 2024. doi: 10.1001/jama.2024.7741.

[14] Liang Dong, Leiyang Chen, Chengliang Zheng, Zhongwang Fu, Umer Zukaib, Xiaohui Cui, and Zhidong Shen. Ocie: Augmenting model interpretability via deconfounded explanation-guided learning. *Knowledge-Based Systems*, 302:112390, 2024. doi: 10.1016/j.knosys.2024.112390.

[15] Z. Fang and Y. Liang. Deep learning approaches for conditional treatment effect estimation. *arXiv preprint arXiv*, 2024.

[16] David M. Fergusson and L. John Horwood. Alcohol abuse and crime: A fixed-effects regression analysis. *Addiction*, 95(10):1525–1536, 2000. doi: 10.1046/j.1360-0443.2000.951015257.x.

[17] Stefan Feuerriegel et al. Causal machine learning. *ACM Computing Surveys*, 2024.

[18] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning, 2022. URL `https://arxiv.org/abs/2212.03954`.

[19] Thomas A. Glass, Steven N. Goodman, Miguel A. Hernán, and Jonathan M. Samet. Causal inference in public health. *Annual Review of Public Health*, 34:61–75, 2013. doi: 10.1146/annurev-publhealth-031811-124606.

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[21] Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350, 2016.

[22] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys*, 53(4), 2023.

[23] Gael P. Hammer, Jean-Baptist du Prel, and Maria Blettner. Avoiding bias in observational studies: Part 8 in a series of articles on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 106(41): 664–668, October 2009. doi: 10.3238/arztebl.2009.0664.

[24] Gemma Hammerton and Marcus R. Munafò. Causal inference with observational data: the need for triangulation of evidence. *Psychological Medicine*, 51(4):563–578, March 2021. doi: 10.1017/S0033291720005127.

[25] Ruth Harrison, Marcus R. Munafò, George Davey Smith, and Robyn E. Wootton. Examining the effect of smoking on suicidal ideation and attempts: triangulation of epidemiological approaches. *The British Journal of Psychiatry*, 217(6):701–707, 2020. doi: 10.1192/bjp.2020.68.

[26] Md Golam Moula Mehedi Hasan and Douglas A. Talbert. Counterfactual examples for data augmentation: A case study. In *Proceedings of the 34th International FLAIRS Conference*, 2021. doi: 10.32473/flairs.v34i1.128503. Using counterfactuals for data augmentation in low-data settings.

[27] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020. URL `https://api.semanticscholar.org/CorpusID:203694352`.

[28] Miguel A. Hernan. *Causal Inference: What If*. Taylor Francis, Boca Raton, 2024.

[29] Guido W. Imbens. *Causal Inference in the Social Sciences*. Cambridge University Press, 2024.

[30] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

[31] Fredrik D Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, 2016.

[32] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[33] Masahiro Kato. Causal-policy forest for end-to-end policy learning, 2025. URL `https://arxiv.org/abs/2512.22846`.

[34] Edward H. Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

[35] Christoph Kern, Michael Kim, and Angela Zhou. Multicate: Multi-accurate conditional average treatment effect estimation robust to unknown covariate shifts, 2024. URL `https://arxiv.org/abs/2405.18206`.

[36] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

[37] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *PNAS*, 2019.

[38] Ruiwen Li, Zhibo Zhang, Jiani Li, Scott Sanner, Jongseong Jang, Yeonjeong Jeong, and Dongsub Shim. Edda: Explanation-driven data augmentation to improve explanation faithfulness. *arXiv preprint arXiv:2105.14162*, 2021. URL `https://arxiv.org/abs/2105.14162`. Explanation-guided data augmentation via occlusion of salient vs non-salient regions.

[39] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL `https://arxiv.org/abs/1705.07874`.

[40] Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

[41] Alessandro Marchese, Jeroen Berrevoets, and Sam Verboven. Causal explanation-guided learning for organ allocation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=b8XmjZmsN0`.

[42] Myrl G. Marmarelis, Greg Ver Steeg, Aram Galstyan, and Fred Morstatter. Ensembled prediction intervals for causal outcomes under hidden confounding. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 18–40. PMLR, 01–03 Apr 2024. URL `https://proceedings.mlr.press/v236/marmarelis24a.html`.

[43] M. Miró-Nicolau, A. Jaume-i Capó, and G. Moyà-Alcover. Role of locality, fidelity and symmetry regularization in learning explainable representations. *Information Sciences*, 2023. explanation priors enforced via regularization without ground-truth explanation labels.

[44] M. Miró-Nicolau, A. Jaume-i-Capó, and G. Moyà-Alcover. Assessing fidelity in xai post-hoc techniques: A comparative study with ground truth explanations datasets. *arXiv preprint arXiv:2311.01961*, 2023.

[45] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 607–617, 2019.

[46] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 2021.

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

[48] PAUL ROSENBAUM and Donald Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70: 41–55, 04 1983. doi: 10.1093/biomet/70.1.41.

[49] Paul R. Rosenbaum. *Design of Observational Studies*, volume 10. Springer, 2010. doi: 10.1007/978-3-030-46405-9.

[50] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

[51] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5): 688–701, 1974. doi: 10.1037/h0037350.

[52] Amal Saadallah. Shap-guided regularization in machine learning models, 2025. URL https://arxiv.org/abs/2507.23665.

[53] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning (ICML)*, pages 3076–3085, 2017.

[54] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

[55] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *NeurIPS*, 2019.

[56] Claudia Shi, David M Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[57] Chiara Sirocchi, Alessandro Bogliolo, and Sara Montagna. Medical-informed machine learning: integrating prior knowledge into medical decision systems. *BMC Medical Informatics and Decision Making*, 24(Suppl 4):186, 2024. doi: 10.1186/s12911-024-02582-4. URL https://doi.org/10.1186/s12911-024-02582-4.

[58] Erik Sverdrup and Yifan Cui. Proximal causal learning of conditional average treatment effects, 2023. URL https://arxiv.org/abs/2301.10913.

[59] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633, 2021. doi: 10.1109/TKDE.2021.3079836.

[60] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 2017.

[61] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests, 2018.

[62] Yong Wu, Yanwei Fu, Shouyan Wang, and Xinwei Sun. Doubly robust proximal causal learning for continuous treatments. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TjGJFkU3xL.

[63] Hao Zhang, Jiayi Chen, Haotian Xue, and Quanshi Zhang. Towards a unified evaluation of explanation methods without ground truth. *arXiv preprint arXiv:1911.09017*, 2019.

[64] Jing Zhang, Yu Jin, and Xin Wang. Comparing meta-learners for estimating heterogeneous treatment effects and conducting sensitivity analyses. *Mathematical and Computational Applications*, 30(6):139, 2025. doi: 10.3390/mca30060139. URL `https://doi.org/10.3390/mca30060139`.

[65] Shichang Zhang, Tessa Han, Usha Bhalla, and Himabindu Lakkaraju. Towards unified attribution in explainable ai, data-centric ai, and mechanistic interpretability, 2025. URL `https://arxiv.org/abs/2501.18887`.