`International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 1, Issue 2 December 2024 ISSN: 3048-975X
https://ijrasht.com/

Causal Machine Learning: Bridging Correlation and Causation in Artificial Intelligence

Palak, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822442.cse@pietgroup.co.in

Riya, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822443.cse@pietgroup.co.in

Asmi, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822425.cse@pietgroup.co.in

Vansh, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822649.cse@pietgroup.co.in

Richa Chaudhary, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India richa.cse-aiml@piet.co.in

Abstract— Most machine learning systems are trained to exploit statistical correlations in data, yet many of the questions practitioners actually care about — what will happen if we intervene, or what would have happened had we acted differently — are inherently causal and cannot be answered from correlation alone. Causal machine learning integrates ideas from the statistical causal inference literature, including the potential outcomes framework and structural causal models, with modern representation learning to estimate cause-and-effect relationships from observational and experimental data. This paper surveys the foundations of causal inference, reviews algorithms for causal discovery and treatment effect estimation, and examines how deep learning has been adapted to produce causally aware representations. We discuss applications in healthcare, economics, recommendation systems, and robotics, summarize commonly used benchmarks, and present a worked example illustrating backdoor adjustment for confounding. We conclude with a discussion of open challenges, including the difficulty of verifying causal assumptions from data alone, and outline promising directions connecting causal machine learning with large-scale representation learning and out-of-distribution generalization.

Index Terms—causal inference, causal machine learning, structural causal models, treatment effect estimation, causal discovery, counterfactual reasoning, out-of-distribution generalization

I. INTRODUCTION

Modern machine learning has achieved impressive predictive accuracy by identifying statistical regularities in large datasets. A model trained to predict hospital readmission, loan default, or product demand typically does so by learning correlations between input features and an outcome variable. This is often sufficient for prediction, but it is not sufficient for decision-making. Deciding whether to prescribe a medication, change a pricing policy, or redesign a recommendation algorithm requires understanding what would happen if an intervention were applied, a fundamentally different question from predicting an outcome under passively observed conditions. As the well-known statistical dictum states, correlation does not imply causation, and models that only capture correlation can mislead decision-makers who intend to act on their predictions.

Causal inference has a long and rigorous history in statistics, epidemiology, and econometrics, formalized through frameworks such as the potential outcomes model developed by Neyman and Rubin, and the structural causal model and do-calculus developed by Pearl and colleagues. These frameworks provide precise mathematical language for defining causal effects, identifying the conditions under which they can be estimated from observational data, and reasoning about confounding, selection bias, and mediation. Until recently, however, this literature developed largely independently of the machine learning community, with each field using different terminology, notation, and, in some cases, differing philosophical commitments about what constitutes a valid causal claim.

Causal machine learning is the emerging synthesis of these two traditions. It applies flexible, high-capacity machine learning models — including deep neural networks, ensemble methods, and representation learning techniques — to the classical problems of causal discovery and causal effect estimation, while retaining the identification guarantees and formal rigor of the statistical causal inference literature. This synthesis has produced practical tools for estimating individualized treatment effects from observational healthcare data, debiasing recommender systems trained on selectively logged user interactions, and building representations that remain valid under distributional shift because they capture causal rather than merely correlational structure.

This paper reviews the state of causal machine learning with the following organization. Section II introduces the conceptual background distinguishing correlation from causation and summarizes the two dominant mathematical frameworks used in the field. Section III presents a taxonomy of causal modeling approaches. Section IV surveys core algorithmic methods for causal discovery and effect estimation, including deep learning based approaches. Section V reviews applications. Section VI summarizes benchmarks. Section VII offers a comparative discussion. Section VIII presents a worked illustrative example. Section IX discusses implementation tooling. Section X considers ethical implications. Section XI discusses open challenges, Section XII outlines future directions, and Section XIII concludes. The growing practical relevance of causal machine learning is also driven by the increasing deployment of automated decision systems in settings where the cost of acting on a spurious correlation is high. In digital advertising, for instance, treating a purely correlational click-prediction model as if it measured the causal effect of an advertisement on purchase behavior can lead to systematically overpaying for impressions that would have converted regardless of whether the advertisement was shown, a problem that has motivated substantial industrial investment in causal measurement techniques. Similar dynamics appear in public policy, where a program evaluation based on correlational analysis of who chose to participate in a voluntary program can substantially overstate the program's true effect, since participants who self-select into a program often differ systematically from non-participants in ways that also affect the outcome being measured, independent of the program itself.

II. BACKGROUND AND MOTIVATION

A. Correlation Versus Causation

A correlation between two variables can arise for several distinct reasons: one variable may cause the other, the relationship may run in the opposite direction, or both variables may be influenced by a third, possibly unobserved, confounding variable. A classical illustration is the historical observation that ice cream sales and drowning incidents rise and fall together; neither causes the other, but both are driven by a common cause, warm weather, that increases both swimming activity and ice cream purchases. A predictive model trained on this data would correctly learn that ice cream sales predict drowning risk, yet acting on this correlation by restricting ice cream sales would have no effect on drowning rates, illustrating why purely predictive models can give catastrophically misleading policy guidance. This distinction matters enormously for machine learning systems deployed to support decisions. A model that predicts loan default well under historical lending patterns may perform poorly, or actively cause harm, if used to decide who receives a loan, because changing lending policy changes the population and the mechanisms generating the outcome, a phenomenon closely related to what economists call the Lucas critique. Causal machine learning aims to build models that remain valid not just for prediction under the observed data-generating process, but for reasoning about the consequences of interventions that change that process.

A further subtlety is that many machine learning benchmarks implicitly reward models for exploiting whatever correlational shortcut is available in the training distribution, even when that shortcut is causally irrelevant to the task the model is nominally meant to perform. A well-documented example from computer vision involves image classifiers that learn to associate certain animals with characteristic backgrounds present in their training images, achieving high benchmark accuracy while relying on a spurious background correlation rather than the causally relevant features of the animal itself, and failing when the same animal appears in an atypical setting at test time. This phenomenon, sometimes termed shortcut learning, illustrates that even purely perceptual tasks can benefit from a causal perspective on which features a model should, in principle, be relying upon.

B. The Potential Outcomes Framework

The potential outcomes framework, associated with the work of Jerzy Neyman and later formalized and popularized by Donald Rubin, defines the causal effect of a treatment for an individual as the difference between the outcome that would occur if the individual received the treatment and the outcome that would occur if the same individual did not receive it. Because any individual is observed under only one of these two conditions at a given time, the other potential outcome is inherently unobserved, a fact often referred to as the fundamental problem of causal inference. Estimating average or individual causal effects therefore requires assumptions — such as confoundedness, meaning that treatment assignment is independent of potential outcomes given observed covariates — that allow the missing potential outcome to be inferred statistically from other individuals with similar covariates.

C. Structural Causal Models and the Causal Hierarchy

Judea Pearl's structural causal model framework represents causal relationships as a directed acyclic graph together with a set of structural equations describing how each variable is generated from its direct causes and an independent noise term. This framework introduces the do-operator, which formally represents the effect of an external intervention that sets a variable to a specific value, overriding its natural causal mechanism, and is distinct from merely conditioning on that variable taking that value in observational data. Pearl further organizes causal reasoning into a three-level hierarchy: associational reasoning, which concerns statistical dependence and is answerable from observational data alone; interventional reasoning, which concerns the effect of hypothetical actions and generally requires either experimental data or additional causal assumptions to answer from observational data; and counterfactual reasoning, which concerns what would have happened in a specific realized scenario had a different action been taken, and which requires the most detailed causal knowledge of the three levels.

This causal hierarchy is a useful organizing principle for understanding the limitations of standard machine learning, which operates almost entirely at the associational level. Because interventional and counterfactual queries generally cannot be answered from associational data without additional assumptions, a central technical concern of causal machine learning is identification: determining, given a causal graph or a set of structural assumptions, whether a desired interventional or counterfactual quantity can in principle be computed from available observational data, and if so, deriving an estimator for it.

III. A TAXONOMY OF CAUSAL MODELING FRAMEWORKS

Causal machine learning methods can be organized according to which of two dominant mathematical frameworks they build upon, and according to which causal task they address. The potential outcomes framework is most naturally suited to treatment effect estimation, where the object of interest is the average or individualized difference in outcome between treatment and control conditions, and where much of the applied literature in economics, epidemiology, and clinical trials is grounded. The structural causal model framework is more naturally suited to representing complex causal systems with many interacting variables, supports reasoning about arbitrary interventions rather than only binary treatment assignment, and provides graphical criteria — such as the backdoor and front-door criteria — for determining when confounding can be adjusted for using observed covariates.

Orthogonal to this distinction between frameworks is a distinction between two causal tasks: causal discovery, which seeks to infer the causal graph structure itself from data, typically using patterns of conditional independence or asymmetries in noise distributions; and causal effect estimation, which assumes the causal structure, or at least the relevant confounding relationships, are already known or hypothesized, and focuses on estimating the magnitude of a specific causal effect. Most practical applications combine elements of both tasks: domain knowledge or discovery algorithms are used to hypothesize a plausible causal structure, which is then used to guide the choice of adjustment variables for effect estimation.

A third, increasingly influential dimension of this taxonomy concerns the source of data used for estimation. Experimental data from randomized controlled trials provides the strongest basis for causal conclusions because randomization ensures, by design, that treatment assignment is independent of all potential outcomes, removing confounding by construction rather than requiring it to be adjusted away statistically. Observational data, by contrast, requires the analyst to identify and adjust for confounding using domain knowledge and statistical technique, and is consequently more susceptible to unverifiable assumptions, but is far more widely available at scale, since large administrative, clinical, and behavioural datasets are collected as a byproduct of ordinary operations rather than requiring a costly and sometimes ethically constrained experimental design. A growing body of causal machine learning research focuses specifically on combining small amounts of experimental data with large amounts of observational data, using the experimental subsample to calibrate or validate assumptions applied to the larger observational dataset.

TABLE I. TAXONOMY OF CAUSAL MACHINE LEARNING FRAMEWORKS AND TASKS

Dimension	Category	Representative Methods	Typical Use Case
Framework	Potential outcomes	Propensity scores, doubly robust estimators, causal forests	Treatment effect estimation from RCT or observational data

Dimension	Category	Representative Methods	Typical Use Case
Framework	Structural causal models	PC/FCI algorithms, do-calculus, backdoor adjustment	Representing multi-variable causal systems, general interventions
Task	Causal discovery	PC, GES, LiNGAM, NOTEARS	Inferring unknown causal graph from data
Task	Effect estimation	TARNet, CEVAE, meta-learners	Estimating known-target causal quantity

IV. CORE METHODS

A. Causal Discovery Algorithms

Constraint-based causal discovery algorithms, such as the PC algorithm, infer a causal graph by testing conditional independence relationships among observed variables and pruning edges accordingly, ultimately producing an equivalence class of graphs consistent with the observed independence structure rather than a single uniquely identified graph, since multiple causal structures can imply the same observational independencies. Score-based methods, such as Greedy Equivalence Search, instead search over the space of possible graphs to maximize a penalized likelihood score, typically also returning an equivalence class rather than a fully oriented graph.

A distinct family of methods achieves full identifiability under additional assumptions about the functional form of the causal mechanisms or the distribution of noise terms. The Linear Non-Gaussian Acyclic Model (LiNGAM), for example, assumes linear structural equations with non-Gaussian noise, which breaks the symmetry that otherwise prevents distinguishing cause from effect using observational data alone, allowing the full causal direction to be recovered rather than only an equivalence class. More recent continuous-optimization approaches, such as NOTEARS, reformulate the combinatorial search over acyclic graphs as a continuous constrained optimization problem, enabling gradient-based methods and, in some variants, deep neural network parameterizations of the structural equations to be used for discovery on larger variable sets.

B. Classical Treatment Effect Estimation

Under the potential outcomes framework, a range of classical estimators address confounding using observed covariates. Propensity score methods estimate the probability of treatment assignment given covariates and use this estimated propensity score for matching, stratification, or inverse-probability weighting, effectively reweighting the observational sample to resemble a randomized experiment. Doubly robust estimators combine an outcome regression model with a propensity score model such that the resulting estimator remains consistent if either, but not necessarily both, of the two component models is correctly specified, providing a useful robustness property against model misspecification that neither approach alone can offer.

C. Deep Learning for Treatment Effect Estimation

Deep learning has been adapted to causal effect estimation primarily by learning a shared representation of covariates that is then used to predict outcomes separately for treated and control conditions. TARNet and its extension, counterfactual regression, learn a representation intended to balance the distribution of treated and control covariates in representation space, reducing the covariate-shift-like discrepancy between treatment groups that would otherwise bias naïve outcome regression. The Causal Effect Variational Autoencoder (CEVAE) instead takes a generative modeling approach, learning a latent variable model of unobserved confounders from observed covariates, treatment, and outcome jointly, under the assumption that the observed covariates act as noisy proxies for the true, unobserved confounder.

Meta-learner approaches, including the S-learner, T-learner, and X-learner, provide a simpler and highly practical alternative by combining off-the-shelf supervised learning models — such as gradient-boosted trees or neural networks — in different ways to estimate heterogeneous treatment effects, without requiring a bespoke architecture specifically designed for causal estimation. Because meta-learners can be implemented using any base learner, they have become a popular practical entry point for applying causal machine learning within existing production machine learning pipelines.

D. Causal Representation Learning and Invariance

A more recent research direction, causal representation learning, seeks to learn representations of raw, high-dimensional data — such as images or text — that correspond to underlying causal variables rather than arbitrary statistical features, motivated by the observation that representations aligned with causal structure should generalize better under distributional shift than representations that merely capture spurious statistical associations present in a particular training distribution. Invariant Risk Minimization (IRM) operationalizes this idea by seeking a representation for which the optimal predictor is the same across multiple observed training environments, on the premise that predictive relationships which hold invariantly across environments are more likely to reflect genuine causal mechanisms rather than environment-specific spurious correlations.

This connection between causal structure and out-of-distribution robustness has become an influential motivation for causal machine learning research beyond classical treatment effect estimation, since it reframes causal learning as a potential solution to the broader and highly practical problem of building models that continue to perform reliably when deployed in environments that differ from their training distribution.

E. Causal Reinforcement Learning

Reinforcement learning agents implicitly reason about the consequences of their own actions, making causal structure directly relevant to the field. Causal reinforcement learning incorporates explicit causal models of the environment to improve sample efficiency, enable more accurate off-policy evaluation of a candidate policy using previously logged data, and support counterfactual reasoning about how an agent's trajectory would have differed under an alternative policy, which is particularly valuable in settings such as healthcare or education where deploying and testing many candidate policies directly in the real environment is costly or unethical.

A particularly active sub-area concerns off-policy evaluation with unobserved confounders, where the logged data used to evaluate a new policy was generated under a previous policy that may have depended on contextual information not recorded in the log, biasing naive importance-sampling estimators of the new policy's expected performance. Causal bandit and causal off-policy evaluation methods explicitly model this confounding structure, using bounds derived from partial causal knowledge to produce conservative but valid performance estimates even when the exact confounding mechanism cannot be fully identified from the available logged data.

V. APPLICATIONS

A. Healthcare and Treatment Effect Estimation

Estimating individualized treatment effects from observational electronic health record data is one of the most mature application areas for causal machine learning, supporting questions such as which patients are likely to benefit most from a given therapy, accounting for confounding introduced by the fact that treatment assignment in observational data is rarely random and often depends on unmeasured clinical judgment.

Causal methods have also been applied to emulate randomized trials using observational data when running an actual trial would be infeasible or unethical, an approach sometimes called the target trial framework, which explicitly specifies the hypothetical randomized experiment being emulated and uses this specification to guide the choice of adjustment variables and estimation strategy applied to the observational data.

B. Economics and Policy Evaluation

Causal inference has long been central to empirical economics, and machine learning based causal methods have extended classical econometric tools to settings with high-dimensional covariates, allowing more flexible modeling of confounding relationships in the estimation of policy effects such as the impact of minimum wage changes, educational interventions, or public health campaigns, while retaining valid statistical inference through techniques such as double or debiased machine learning that explicitly correct for the bias introduced by using flexible machine learning models for nuisance parameter estimation.

Machine learning has also extended quasi-experimental econometric designs, such as regression discontinuity and difference-in-differences estimators, by allowing flexible, nonparametric modeling of trends and covariate relationships that would traditionally have been restricted to simple linear specifications, while generalized random forests and related tree-based ensemble methods have been used to estimate how policy effects vary systematically across subpopulations, informing more targeted rather than uniformly applied policy interventions.

C. Recommender Systems and Debiasing

Recommender systems trained on logged user interaction data face a subtle causal challenge: the items a user interacted with were themselves selected by a previous recommendation policy, creating a feedback loop in which popular items receive more exposure, more interactions, and therefore more training signal, regardless of their true relevance to users. Causal debiasing techniques model this exposure mechanism explicitly and apply inverse-propensity weighting or related causal adjustment techniques to correct for this selection bias, improving fairness and diversity of recommendations relative to naively trained systems.

Related causal techniques are used in digital advertising to estimate the incremental effect of showing an advertisement to a particular user, as distinct from simply predicting the probability that the user converts, since a substantial fraction of users who convert after seeing an advertisement would have converted regardless, and correctly identifying this incremental, causally attributable effect is central to efficient advertising budget allocation.

D. Explainable AI via Causal Reasoning

Causal methods provide an alternative foundation for explainability distinct from purely correlational feature-importance techniques. Causal explanations aim to answer the counterfactual question of how a model's prediction would change had a specific input feature been different, holding other causally relevant factors fixed according to the underlying causal graph, which can produce explanations that are more actionable and less misleading than standard feature attribution methods that do not account for causal dependencies among input features.

This causal framing is particularly relevant for recourse-oriented explanations, which aim to tell an individual affected by an automated decision what they could feasibly change to obtain a different outcome in the future. A purely correlational recourse recommendation might suggest changing a feature that is itself a consequence rather than a cause of the outcome, or that is causally entangled with immutable characteristics, producing recourse advice that is either infeasible or misleading, whereas a causally grounded recourse method restricts suggested changes to features that are genuinely manipulable and causally upstream of the decision.

E. Robotics and Causal World Models

In robotics, causal world models allow an agent to reason about the effect of its actions on the environment in a way that generalizes to novel situations not directly observed during training, supporting more sample-efficient planning than purely model-free reinforcement learning, particularly in settings where the physical consequences of an action are governed by stable, reusable causal mechanisms, such as rigid-body dynamics, that remain valid even as surface-level visual appearance changes.

Causal world models are also being explored for sim-to-real transfer, where an agent trained primarily in simulation must generalize to a physical robot operating in a real environment; by explicitly separating causal mechanisms that transfer reliably between simulation and reality, such as basic physical dynamics, from surface-level visual or sensor characteristics that do not transfer, causally structured world models aim to reduce the performance gap that has historically limited direct transfer of simulation-trained policies to physical hardware.

VI. BENCHMARKS AND EVALUATION

Evaluating causal machine learning methods is inherently more difficult than evaluating standard predictive models, because the ground-truth causal effect is by definition unobserved in real data, a direct consequence of the fundamental problem of causal inference described in Section II. The field has therefore relied heavily on semi-synthetic benchmarks, in which real covariates are combined with a simulated treatment assignment and outcome-generating mechanism whose true causal effect is known by construction, allowing estimator accuracy to be measured directly against ground truth. Table II summarizes several widely used benchmarks for treatment effect estimation and causal discovery.

The IHDP benchmark, derived from a real infant health and development randomized study with a semi-synthetic outcome model, and the Jobs benchmark, combining data from a real job training program with experimental and observational subsamples, are among the most widely reported benchmarks for individualized treatment effect estimation, typically evaluated using the precision in estimation of heterogeneous effects metric. For causal discovery, the Sachs protein-signaling dataset, derived from real flow cytometry measurements with an expert-validated ground-truth causal graph, remains a standard reference point despite its relatively small size, precisely because few real-world datasets offer independently validated ground-truth causal structure against which discovery algorithms can be benchmarked.

A recurring methodological concern with semi-synthetic benchmarks is that the simulated component of the data-generating process, typically the outcome model or the treatment assignment mechanism, is chosen by the benchmark designer and may not reflect the complexity or irregularity of real-world confounding relationships, meaning strong performance on such a benchmark does not guarantee comparable performance in a genuinely novel observational dataset with different, unknown confounding structure. This has led some researchers to advocate for a portfolio evaluation approach, in which a proposed method is tested across many distinct semi-synthetic benchmarks with varying confounding strength, functional form, and dimensionality, on the premise that consistent performance across a diverse benchmark portfolio provides more convincing evidence of real-world robustness than strong performance on any single benchmark.

TABLE II. REPRESENTATIVE BENCHMARKS IN CAUSAL MACHINE LEARNING

Benchmark	Task	Data Basis	Ground Truth Source
IHDP	Treatment effect estimation	Real covariates, simulated outcome	Known simulation parameters
Jobs	Treatment effect estimation	Real experimental + observational data	Randomized subsample
Twins	Treatment effect estimation	Real twin birth records	Twin pair as counterfactual proxy
Sachs	Causal discovery	Real protein flow cytometry data	Expert-validated biological pathway
SynTReN	Causal discovery	Simulated gene regulatory networks	Known simulation graph

VII. COMPARATIVE ANALYSIS AND DISCUSSION

Classical potential-outcomes-based estimators such as propensity score methods remain attractive for their statistical simplicity, interpretability, and well-understood asymptotic properties, but they typically assume a relatively low-dimensional set of confounders and can perform poorly when confounding is driven by high-dimensional or unstructured data such as images or free text. Deep-learning-based estimators such as TARNet and CEVAE extend causal estimation to such high-dimensional settings, at the cost of losing some of the statistical guarantees and interpretability of classical methods, and generally requiring larger sample sizes to train reliably. Meta-learner approaches occupy a useful middle ground, allowing flexible, high-capacity base learners while retaining a comparatively simple and interpretable overall estimation structure.

Causal discovery methods face a distinct trade-off between assumption strength and identifiability. Constraint-based and score-based methods make comparatively weak assumptions but generally cannot fully orient every edge in the causal graph, returning an equivalence class rather than a unique answer. Methods that assume specific functional forms, such as LiNGAM, can achieve full identifiability but are only valid when their functional and distributional assumptions genuinely hold in the data-generating process, and violations of these assumptions can lead to confidently wrong causal conclusions rather than a graceful degradation in accuracy, which is an important practical caveat for practitioners applying these methods outside carefully controlled settings.

A practical consequence of these trade-offs is that experienced practitioners rarely commit to a single method in isolation. A common workflow begins with constraint-based or score-based discovery to narrow the space of plausible causal graphs using comparatively weak, broadly applicable assumptions, followed by domain-expert review to select among the remaining equivalence class members, and concludes with effect estimation using a doubly robust or meta-learner approach chosen for its robustness properties given the confounding structure implied by the selected graph. This layered workflow reflects a broader principle in causal machine learning: because no single method's assumptions can be fully verified from data, methodological triangulation across multiple approaches with different assumption profiles, combined with domain expertise, generally provides more trustworthy conclusions than reliance on any individual technique.

TABLE III. COMPARATIVE SUMMARY OF CAUSAL EFFECT ESTIMATION APPROACHES

Approach	Confounder Dimensionality	Interpretability	Statistical Guarantees
Propensity score methods	Low to moderate	High	Well-established asymptotics
Doubly robust estimators	Moderate	High	Robust to single model misspecification
Meta-learners	Moderate to high	Moderate	Depends on base learner
Deep representation methods (TARNet, CEVAE)	High / unstructured	Lower	Fewer formal guarantees

VIII. ILLUSTRATIVE CASE STUDY

Consider a simplified but representative scenario: a hospital wants to estimate the effect of a new medication on patient recovery time using observational records, where physicians were more likely to prescribe the medication to patients they judged, based on unrecorded clinical intuition partially correlated with recorded severity scores, to be more likely to benefit. A naïve comparison of average recovery time between medicated and unmedicated patients would be confounded by this severity-related prescribing pattern, since more severe patients who received the medication might recover more slowly on average not because the medication is harmful, but because they started from a worse baseline condition.

Applying the backdoor adjustment criterion from the structural causal model framework, if recorded severity scores fully capture the confounding influence of clinical judgment on both prescribing decisions and recovery outcomes, the causal effect of the medication can be estimated by stratifying patients into severity strata, computing the treated-versus-untreated recovery time difference within each stratum, and averaging these within-stratum effects across the severity distribution, effectively comparing patients who are similar in their measured severity but differ in whether they received the medication. This adjustment removes the confounding bias introduced by severity, provided the confoundedness assumption holds, namely that no additional unmeasured factor jointly influences both prescribing and recovery beyond what is captured by the recorded severity score.

In practice, this adjustment is implemented using either explicit stratification, propensity-score-based reweighting, or a doubly robust combination of an outcome regression model and a propensity model, as described in Section IV. The critical and often underappreciated point illustrated by this example is that the validity of the resulting causal estimate depends entirely on an assumption — that severity scores capture all relevant confounding — that cannot itself be verified from the observational data alone, and must instead be justified using clinical domain knowledge, underscoring why causal machine learning practitioners must combine statistical technique with careful subject-matter judgment rather than treating causal estimation as a purely automated procedure.

A natural extension of this example illustrates why sensitivity analysis is a standard complement to point estimation in applied causal machine learning. Suppose a clinician suspects that some additional factor, such as a patient's home support situation, influences both prescribing decisions and recovery but is not recorded in the dataset. Rather than abandoning the analysis, a sensitivity analysis quantifies how strong such an unmeasured confounder would need to be, in terms of its association with both treatment and outcome, to fully explain away the estimated effect. If the estimated effect would only be nullified by an implausibly strong unmeasured confounder relative to the strength of measured confounders such as severity, the estimate is considered comparatively robust; if a plausible unmeasured confounder could easily account for the entire estimated effect, the finding should be treated with substantially more caution. This style of reasoning allows practitioners to communicate a defensible degree of confidence in a causal conclusion even when the strong confoundedness assumption cannot be verified outright.

IX. IMPLEMENTATION CONSIDERATIONS AND TOOLING

A. Software Libraries

Several open-source libraries have substantially lowered the barrier to applying causal machine learning in practice. Libraries built around the potential outcomes framework provide implementations of meta-learners, doubly robust estimators, and causal forests with interfaces resembling standard supervised learning APIs, allowing practitioners familiar with conventional machine learning workflows to apply causal estimators with comparatively little additional conceptual overhead. Separate libraries focused on structural causal models provide tools for specifying a causal graph, checking whether a desired causal query is identifiable given that graph using automated do-calculus derivations, and performing sensitivity analysis to quantify how robust an estimate is to potential violations of the confoundedness assumption.

B. Practical Deployment Challenges

Deploying causal machine learning in production settings introduces challenges beyond estimator selection. Causal graphs and adjustment sets are typically specified by domain experts and must be revisited as understanding of the underlying domain evolves or as new confounding pathways are identified, making causal models a comparatively higher-maintenance artifact than a purely predictive model whose validity depends only on the stability of the input-output statistical relationship. Organizations adopting causal machine learning at scale increasingly pair automated estimation pipelines with structured processes for eliciting, documenting, and periodically reviewing the causal assumptions underlying each deployed model.

C. Monitoring and Model Governance

Because causal effect estimates cannot be validated against a held-out ground truth in the way predictive accuracy can, monitoring a deployed causal model requires different practices than monitoring a standard predictive model. Common approaches include periodically re-running discovery or sensitivity analyses as new data accumulates, tracking whether the estimated effect remains stable across successive data refreshes as a rough proxy for structural stability, and, where ethically and practically feasible, periodically validating a subset of causal conclusions against small randomized experiments to check that the observational estimate remains consistent with a directly randomized benchmark. This practice of periodically calibrating observational estimates against experimental data is sometimes referred to as a causal audit and is increasingly viewed as a governance best practice for high-stakes deployments of causal machine learning, analogous to periodic recalibration checks used for predictive models subject to concept drift.

X. ETHICAL AND SOCIETAL CONSIDERATIONS

Causal machine learning is frequently motivated by fairness concerns, since many notions of algorithmic fairness are most naturally expressed as causal rather than purely statistical criteria. Counterfactual fairness, for example, requires that a model's prediction for an individual would remain unchanged had that individual belonged to a different demographic group, holding all non-descendant causal factors fixed, a criterion that cannot be expressed using correlational fairness metrics alone and that depends on an assumed causal graph relating demographic attributes to other observed features.

At the same time, the reliance of causal methods on an assumed causal graph introduces its own ethical risk: an incorrectly specified causal graph can produce a fairness intervention that appears principled while in fact enforcing an unjustified or even harmful assumption about the causal role of a demographic attribute. Because causal assumptions are generally not fully verifiable from data, organizations deploying causal fairness interventions bear a particular responsibility to document and justify the causal assumptions underlying such interventions, ideally with input from domain experts and affected stakeholders, rather than treating a chosen causal graph as a purely technical implementation detail.

A further ethical consideration concerns the use of causal machine learning for targeted intervention at the individual level, such as identifying which specific patients or customers are most responsive to a particular treatment or offer. While this individualized targeting can improve overall efficiency of an intervention, it also raises distributive justice questions about who is included in or excluded from beneficial interventions on the basis of a predicted causal effect, particularly when the underlying estimate is less reliable for subpopulations that were underrepresented in the training data, a concern that mirrors well-documented fairness issues in purely predictive models but takes on additional weight when the resulting decision is justified specifically on causal grounds.

XI. CHALLENGES AND OPEN PROBLEMS

The most fundamental challenge in causal machine learning is that causal assumptions, such as confoundedness or the correctness of a hypothesized causal graph, are generally not testable using observational data alone, meaning that the validity of any causal estimate ultimately rests on domain knowledge and judgment that lies outside the statistical procedure itself. This is a qualitatively different situation from standard supervised learning, where held-out test data can directly validate predictive accuracy, and it means that causal machine learning results should generally be reported and interpreted together with an explicit statement of the assumptions relied upon and, where possible, a sensitivity analysis characterizing how conclusions would change under plausible violations of those assumptions.

A second challenge concerns causal discovery in the presence of latent confounders, cyclic feedback relationships, or non-stationary data-generating processes, all of which are common in real-world systems such as biological networks or economic markets but violate the acyclicity and causal sufficiency assumptions underlying many standard discovery algorithms. A third challenge is scalability: exact structural learning and identification procedures can scale poorly to the very high-dimensional variable sets found in genomics or large-scale recommendation systems, motivating continued research into approximate and continuous-optimization-based discovery methods.

Finally, the causal machine learning literature has historically lacked large-scale, real-world benchmarks with independently verified ground-truth causal effects, relying instead on semi-synthetic constructions whose realism is difficult to fully assess, which complicates rigorous comparison across newly proposed methods and makes it difficult to know how much of a given benchmark improvement will translate to genuine gains in real deployment settings.

A further, often underappreciated challenge concerns communicating causal uncertainty to non-technical decision-makers. A causal effect estimate produced by a sophisticated machine learning pipeline can appear as authoritative and precise as any other model output, obscuring the fact that its validity rests on assumptions that domain experts, not the estimation algorithm itself, must ultimately judge to be plausible. Effective use of causal machine learning in practice therefore depends as much on organizational processes for eliciting and scrutinizing causal assumptions as on the statistical sophistication of the estimation method itself, a socio-technical dimension of the field that has received comparatively less research attention than algorithmic development.

XII. FUTURE RESEARCH DIRECTIONS

One promising direction integrates causal representation learning with large-scale foundation models, seeking to learn representations from massive unlabelled datasets that capture disentangled causal factors of variation rather than arbitrary statistical features, which could substantially improve the out-of-distribution robustness and transferability of pretrained representations across domains and tasks. A second direction explores combining large language models with formal causal reasoning engines, using the language model to extract candidate causal hypotheses or relevant domain knowledge from text, while a formal causal inference engine verifies identifiability and estimates the resulting effect, echoing the tool-augmented reasoning patterns discussed in related neuro-symbolic AI research.

A further direction concerns causal inference under continuous data streams and non-stationary environments, extending classical causal discovery and estimation techniques, which typically assume a fixed underlying causal structure, to settings where causal mechanisms themselves may drift over time, as is common in domains such as online advertising or evolving biological systems. Finally, there is growing interest in developing standardized reporting practices for causal machine learning studies, analogous to established reporting guidelines in clinical epidemiology, that would require explicit documentation of assumed causal graphs, identification strategies, and sensitivity analyses, improving the reproducibility and trustworthiness of causal claims made using machine learning methods.

Another emerging direction involves multi-modal causal representation learning, which seeks to identify shared causal factors underlying data collected from different modalities, such as images, sensor readings, and text describing the same underlying physical or social process. If successful, such methods could allow causal knowledge learned from one modality, for example a well-instrumented simulation environment, to transfer to another modality, such as sparse real-world sensor data, by aligning representations at the level of shared causal variables rather than surface-level statistical features, a capability that would be particularly valuable in scientific and industrial domains where richly instrumented data is available only in controlled settings but decisions must ultimately be made using sparser real-world observations.

XIII. CONCLUSION

Causal machine learning addresses a fundamental limitation of purely correlational predictive models by providing formal tools for reasoning about interventions and counterfactuals, questions that are central to using machine learning to support real-world decisions rather than passive prediction alone. This survey reviewed the potential outcomes and structural causal model frameworks, summarized methods for causal discovery and treatment effect estimation including recent deep learning based approaches, and examined applications spanning healthcare, economics, recommendation systems, explainability, and robotics. While the field faces significant challenges related to the unverifiability of causal assumptions from data alone, scalability of discovery methods, and the scarcity of real-world benchmarks with verified ground truth, its growing integration with representation learning and large-scale pretrained models suggests that causal reasoning will play an increasingly central role in building machine learning systems that are robust, fair, and suitable for informing genuine decisions rather than only generating predictions.

Looking ahead, the maturation of causal machine learning as a discipline is likely to depend as much on the development of shared standards for documenting and communicating causal assumptions as on further algorithmic innovation. Just as the broader machine learning community has converged on common practices for reporting predictive model performance, such as standardized train-test splits and cross-validation protocols, the causal machine learning community would benefit from analogous conventions for reporting the causal graph, identification strategy, and sensitivity analysis underlying a given causal claim. Establishing such conventions would make it considerably easier for practitioners, reviewers, and downstream decision-makers to assess the credibility of a causal machine learning result, and would help the field build the kind of cumulative, trustworthy evidence base that has proven valuable in more established causal disciplines such as clinical epidemiology.

REFERENCES

[1] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[2] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," Journal of Educational Psychology, vol. 66, no. 5, 1974.
[3] J. Neyman, "On the application of probability theory to agricultural experiments," Statistical Science, vol. 5, 1990 (translated from 1923 original).
[4] P. Spirtes, C. Glymour, and R. Scheines, Causation, Prediction, and Search, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.
[5] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," Journal of Machine Learning Research, vol. 7, 2006.
[6] D. M. Chickering, "Optimal structure identification with greedy search," Journal of Machine Learning Research, vol. 3, 2002.
[7] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," in Proc. NeurIPS, 2018.
[8] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in Proc. ICML, 2017.
[9] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in Proc. NeurIPS, 2017.

[10] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," Journal of the American Statistical Association, vol. 113, 2018.
[11] S. Künzel, J. Sekhon, P. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," Proceedings of the National Academy of Sciences, vol. 116, 2019.
[12] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," The Econometrics Journal, vol. 21, 2018.
[13] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," arXiv preprint, 2019.
[14] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," Proceedings of the IEEE, vol. 109, 2021.
[15] J. Peters, D. Janzing, and B. Schölkopf, Elements of Causal Inference: Foundations and Learning Algorithms. Cambridge, MA, USA: MIT Press, 2017.
[16] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," Proceedings of the National Academy of Sciences, vol. 113, 2016.
[17] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," Proceedings of the National Academy of Sciences, vol. 113, 2016.
[18] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in Proc. NeurIPS, 2017.
[19] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," ACM Computing Surveys, vol. 53, 2020.
[20] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," ACM Transactions on Knowledge Discovery from Data, vol. 15, 2021.
[21] J. M. Robins, M. A. Hernan, and B. Brumback, "Marginal structural models and causal inference in epidemiology," Epidemiology, vol. 11, 2000.
[22] M. A. Hernan and J. M. Robins, "Using big data to emulate a target trial when a randomized trial is not available," American Journal of Epidemiology, vol. 183, 2016.
[23] N. Kallus, "Balanced policy evaluation and learning," in Proc. NeurIPS, 2018.
[24] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," in Proc. ICML, 2016.
[25] Y. Saito and T. Joachims, "Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances," in Proc. RecSys, 2021.
[26] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising," Journal of Machine Learning Research, vol. 14, 2013.
[27] J. Buesing, T. Weber, Y. Zwols, S. Racaniere, A. Guez, J.-B. Lespiau, and N. Heess, "Woulda, coulda, shoulda: Counterfactually-guided policy search," in Proc. ICLR, 2019.
[28] E. Bareinboim, A. Forney, and J. Pearl, "Bandits with unobserved confounders: A causal approach," in Proc. NeurIPS, 2015.
[29] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," arXiv preprint, 2020.
[30] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath, "CausalGAN: Learning causal implicit generative models with adversarial training," in Proc. ICLR, 2018.
[31] P. Schölkopf, "Causality for machine learning," arXiv preprint, 2019.
[32] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," Science, vol. 308, 2005.
[33] A. Hill, "Bayesian nonparametric modeling for causal inference," Journal of Computational and Graphical Statistics, vol. 20, 2011.
[34] N. Hassanpour and R. Greiner, "Learning disentangled representations for counterfactual regression," in Proc. ICLR, 2020.
[35] F. D. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in Proc. ICML, 2016.
[36] A. Curth and M. van der Schaar, "Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms," in Proc. AISTATS, 2021.
[37] D. Bahadori, K. Heckerman, and V. Sun, "Causal regularization," in Proc. NeurIPS, 2017.
[38] C. Heinze-Deml, J. Peters, and N. Meinshausen, "Invariant causal prediction for nonlinear models," Journal of Causal Inference, vol. 6, 2018.
[39] V. Chernozhukov, C. Hansen, and M. Spindler, "Post-selection and post-regularization inference in linear models with many controls and instruments," American Economic Review, vol. 105, 2015.
[40] J. D. Angrist and J.-S. Pischke, Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ, USA: Princeton Univ. Press, 2009.
[41] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," The Annals of Statistics, vol. 47, 2019.
[42] R. Moraffah, P. Sheth, M. Karami, A. Bhattacharya, Q. Wang, A. Tahir, A. Raglin, and H. Liu, "Causal inference for time series analysis: Problems, methods and evaluation," Knowledge and Information Systems, vol. 63, 2021.
[43] P. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, "Causal inference and counterfactual prediction in machine learning for actionable healthcare," Nature Machine Intelligence, vol. 2, 2020.
[44] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, E. H. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler, "Inferring causation from time series in Earth system sciences," Nature Communications, vol. 10, 2019.