A Review of Causal Decision Making

LIN GE*, Amazon, USA
HENGRUI CAI*, University of California, Irvine, USA
RUNZHE WAN*, Amazon, USA
YANG XU*, North Carolina State University, USA
RUI SONG†, Amazon, USA

To make effective decisions, it is important to have a thorough understanding of the causal relationships among actions, environments, and outcomes. This review aims to surface three crucial aspects of decision making through a causal lens: 1) the discovery of causal relationships through causal structure learning, 2) understanding the impacts of these relationships through causal effect learning, and 3) applying the knowledge gained from the first two aspects to support decision making via causal policy learning. Moreover, we identify challenges that hinder the broader utilization of causal decision making and discuss recent advances in overcoming these challenges. Finally, we provide future research directions to address these challenges and further enhance the implementation of causal decision making in practice, with real-world applications illustrated through the proposed causal decision-making workflow. To facilitate broader adoption, we additionally integrate relevant methods into a unified Python-based collection, offering a methodological and practical framework for the community (available at https://causaldm.github.io/Causal-Decision-Making).

JAIR Track: Surveys

JAIR Associate Editor: Alessandro Farinelli

JAIR Reference Format:
Lin Ge*, Hengrui Cai*, Runzhe Wan*, Yang Xu*, and Rui Song†. 2026. A Review of Causal Decision Making. *Journal of Artificial Intelligence Research* 85, Article 41 (April 2026), 64 pages. DOI: 10.1613/jair.1.21001

1 Introduction

Decision making is at the heart of artificial intelligence systems, enabling agents to navigate complex environments, achieve goals, and adapt to changing conditions. Traditional decision-making frameworks often rely on associations or statistical correlations between variables, which can lead to suboptimal outcomes when the underlying causal relationships are ignored (Pearl 2009). The rise of causal inference as a field has provided powerful frameworks and tools to address these challenges, such as structural causal models and potential outcome frameworks (Pearl 2000; Rubin 1978). Unlike traditional methods, *causal decision making* focuses on identifying and leveraging cause-effect relationships, allowing agents to reason about the consequences of their actions, predict counterfactual scenarios, and optimize decisions in a principled way (Spirtes, C. N. Glymour, et al. 2000). In recent years, numerous decision-making methods based on causal reasoning have been developed, with

*Equal contribution.
†Corresponding author.

Authors' Contact Information: Lin Ge*, ORCID: 0009-0007-3948-9535, gelin9708@gmail.com, Amazon, Seattle, USA; Hengrui Cai*, ORCID: 0000-0002-5679-8862, hengrc1@uci.edu, University of California, Irvine, Irvine, USA; Runzhe Wan*, ORCID: 0009-0000-7820-4271, runzhe.wan@gmail.com, Amazon, New York, USA; Yang Xu*, ORCID: 0009-0008-6473-8529, yxu63@ncsu.edu, North Carolina State University, Raleigh, USA; Rui Song†, ORCID: 0000-0003-1875-2115, songray@gmail.com, Amazon, Seattle, USA.

DOI: 10.1613/jair.1.21001

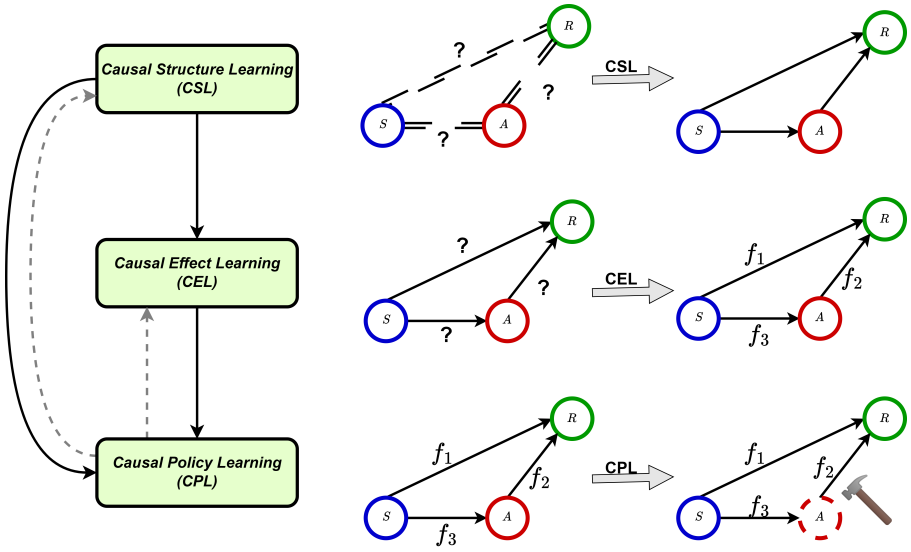


Fig. 1. The workflow of Causal Decision Making. f_1, f_2, and f_3 represent the impact sizes of the directed edges. Variables enclosed in solid circles are observed, while those in dashed circles are actionable.

applications spanning recommender systems (Q. Zhou et al. 2017), clinical trials (Durand et al. 2018), finance (Bai et al. 2024), and ride-sharing platforms (Wan, S. Zhang, et al. 2021). Despite these advancements, a fundamental question persists:

When and why do we need causal modeling in decision making?

This question is closely tied to the concept of counterfactual thinking, which examines what might have happened under alternative decisions or actions. Counterfactual analysis is crucial in domains where the outcomes of unchosen decisions are challenging, if not impossible, to observe. For example, a business leader who selects one marketing strategy over another may never fully know the outcome of the unselected option (Pearl 2009; Rubin 1974). Similarly, in econometrics, epidemiology, psychology, and social sciences, *the inability to observe counterfactuals directly often requires causal approaches* (G. W. Imbens and Rubin 2015; Morgan and Winship 2015). Conversely, non-causal analysis may suffice in scenarios where alternative outcomes are readily determinable. For example, a personal investor's actions may have a negligible impact on the dynamics of the stock market, allowing the potential outcomes of alternative investment decisions to be inferred from existing time series of stock price (Angrist and Pischke 2008). However, even in cases where counterfactual outcomes are theoretically calculable—such as in environments with known models like AlphaGo-exhausting the computation of all possible outcomes remains computationally infeasible (Silver, Hubert, et al. 2018; Silver, Schrittwieser, et al. 2017). In such scenarios, causal modeling remains advantageous, providing *structured ways to infer outcomes efficiently and make robust decisions.*

Most existing studies assume either sophisticated prior knowledge or strong causal models to facilitate follow-up decision making. To make effective and reliable decisions, it is essential to have a thorough understanding of the causal relationships among actions, environments, and outcomes. This review synthesizes the current state of research in Causal Decision Making (CDM), providing an overview of its fundamental concepts, recent advances, and practical applications. Specifically, this work discusses the connections among **three primary components of decision making** through a causal lens: 1) discovering causal relationships via *Causal Structure Learning*

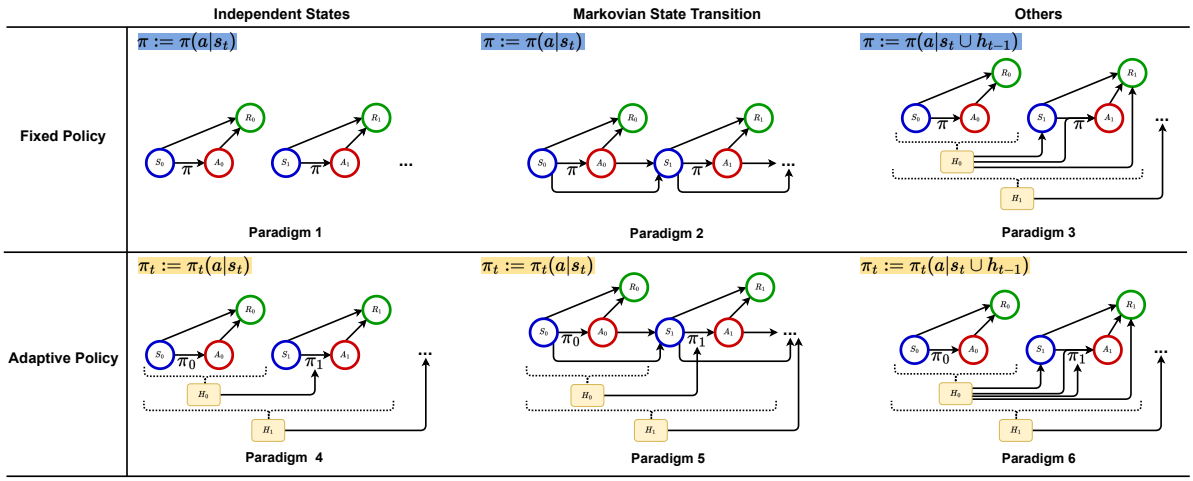


Fig. 2. Common data dependence structures (paradigms) in CDM. Detailed notations and explanations can be found in Section 4.2.

(CSL), 2) understanding the impacts of these relationships through *Causal Effect Learning (CEL)*, and 3) applying the knowledge derived from the first two components to guide decision making via *Causal Policy Learning (CPL)*.

Let S denote the state of the environment, which includes all relevant feature information about the environment the decision-makers interact with, A the action taken, π the action policy that determines which action to take, and R the reward observed after taking action A. As illustrated in Figure 1, CDM typically begins with CSL, which aims to uncover the unknown causal relationships among various variables of interest. Once the causal structure is established, CEL is used to assess the impact of a specific action on the outcome rewards. To further explore more complex action policies and refine decision-making strategies, CPL is used to evaluate a given policy or identify an optimal policy. In practice, it is also common to move directly from CSL to CPL without conducting CEL. Furthermore, CPL has the potential to improve both CEL and CSL by facilitating the development of more efficient experimental design (Simchi-Levi and C. Wang 2023; S. Zhu et al. 2020) and adaptive refinement of causal structures (Sauter et al. 2024).

Building on this framework, decision-making problems discussed in the literature can be further categorized into **six paradigms**, as summarized in Figure 2. These paradigms summarize the common assumptions about data dependencies frequently employed in practice. Paradigms 1-3 describe the data structures in offline learning settings, where data are collected according to an unknown but fixed behavior policy. In contrast, paradigms 4-6 capture online learning settings, where policies dynamically adapt to newly collected data, enabling continuous policy improvement. These paradigms also reflect different assumptions about state dependencies. The simplest cases, i.e., paradigms 1 and 4, assume that all observations are independent, implying no long-term effects of actions on future observations. To account for sequential dependencies, the Markov Decision Process (MDP) framework, summarized in paradigms 2 and 5, assumes that what happens next depends only on the current state and action. When such independence assumptions do not hold, paradigms 3 and 6 account for scenarios where all historical observations may influence state transitions and rewards. This includes, but is not limited to, research on Partially Observable Markov Decision Process (POMDP) (Hausknecht and Stone 2015; Littman 2009), panel data analysis (Hsiao 2022; Hsiao 2007), and Dynamic Treatment Regimes (DTR) with finite stages (Chakraborty and Moodie 2013; Chakraborty and Murphy 2014).

Each CDM task has been studied under different paradigms, with CSL being extensively studied within paradigm 1. CEL and offline CPL cover paradigms 1-3, while online CPL covers paradigms 4-6. By structuring the discussion around these three tasks and six paradigms, this review aims to provide a cohesive framework for understanding the field of Causal Decision Making across diverse tasks and data settings.

Contribution. In this paper, we conduct a comprehensive survey of CDM. Our contributions are as follows.

- We present the first unified framework for causal decision making (CDM) by introducing three core tasks and six paradigms. This framework bridges topics across previously disconnected fields, including economics, statistics, machine learning, and reinforcement learning, all of which include research related to CDM.
- For each task and paradigm, we introduce key concepts that connect diverse literatures and provide a structured review of recent developments. This comprehensive overview of CDM addresses gaps in prior surveys, which often focus narrowly on specific tasks or paradigms and overlook the broader connection between causality and decision-making (see Section 2).
- We provide real-world examples to illustrate the critical role of causality in decision making and to demonstrate how CSL, CEL, and CPL are inherently interconnected in everyday applications—often without being explicitly recognized.
- We are actively maintaining and expanding a GitHub repository and accompanying online book that provide detailed explanations of the key methods reviewed in this paper, together with a code package and demonstration examples to support their implementation (available at: https://causaldm.github.io/Causal-Decision-Making).

The remainder of this paper is organized as follows. Section 2 provides an overview of related survey papers. Section 3 introduces the fundamental concepts, assumptions, and notations that underpin the subsequent discussions. In Section 4, we present a detailed introduction to the three key tasks and six learning paradigms in CDM. Sections 5 through 8 constitute the core of the paper, with each section dedicated to a specific topic within CDM: CSL, CEL, Offline CPL, and Online CPL, respectively. Section 9 then explores the extensions needed when standard causal assumptions are violated. To illustrate the practical application of the CDM framework, Section 10 presents two real-world case studies. Finally, Section 11 discusses additional open research directions that extend beyond those covered earlier, and Section 12 concludes the paper with a summary of our main contributions.

2 Related Work

Many reviews have examined causality or decision making. However, to the best of our knowledge, they typically focus on a single paradigm or task, with some emphasizing methodologies without clearly delineating the fundamental connections between causality and decision making. In contrast, this review proposes a unified framework that integrates all key components, explicitly illustrating the role of causality and the interrelationships across different stages of the decision-making process. A detailed discussion of related surveys follows.

2.1 Causal Inference and Causal Structural Learning

In recent years, several review papers have emerged in the field of causal inference, typically categorized into two main frameworks: the Structural Equation Models (SEM) framework introduced by Pearl (Pearl 1995) and the potential outcomes framework pioneered by Rubin (Rubin 1978, 1974).

The SEM framework models causal relationships using graphical structures, where nodes represent variables and directed edges depict cause-effect relationships. Pearl's work (Pearl 2003) was instrumental in developing the do-calculus to formalize the effects of interventions on causal diagrams. Subsequent reviews (Pearl 2010a, 2009, 2010b) provide comprehensive overviews of CSL, with detailed discussions of related topics such as confounding

issues and mediation analysis. To the best of our knowledge, no recent reviews have comprehensively summarized the latest advances in causal discovery within the SEM framework, which is a key focus of Section 5.

The potential outcomes framework, also known as the Rubin Causal Model (RCM) (Rubin 1974), defines causal effects by comparing potential outcomes under different treatment conditions. A key resource in this area is G. W. Imbens and Rubin (2015), which systematically summarizes the origins and development of causal inference under the RCM, covering topics ranging from estimation and inference to sensitivity analysis. Recent reviews have also explored specific aspects of causal inference, including observational studies (Yao et al. 2021), matching methods (Stuart 2010), missing data (Ding and F. Li 2018), techniques in recommendation systems (Gao et al. 2024), and confounding in text analysis (K. Keith et al. 2020). This part of the work aligns closely with CEL, where existing reviews often fail to summarize effect learning comprehensively across different data structures and paradigms. Our work addresses this gap by reviewing studies under both assumption-satisfying and assumption-violating scenarios and positioning CEL as a critical intermediate stage in the decision-making process.

2.2 Policy Learning

In the area of offline policy learning, related review articles can be classified as focusing on Off-Policy Evaluation (OPE) and Off-Policy Optimization (OPO). For OPE, Voloshin et al. (2019) systematically studies the empirical performance of a list of common OPE methods for the offline Reinforcement Learning (RL) setting (i.e., paradigm 2). Uehara, Shi, et al. (2022) is the latest review of the key methods and theories in OPE, covering paradigms 1-3. For OPO, Prudencio et al. (2023) and Levine, Kumar, et al. (2020) both review key concepts, methods, and open problems in offline RL (paradigm 2). Besides, from a statistical perspective, M. R. Kosorok and Laber (2019) comprehensively reviewed the progress of applying DTR to precision medicine, covering paradigms 1 and 3.

In contrast, most reviews on online policy learning focus on policy optimization, with online policy evaluation being a newer and less explored area. For policy optimization in paradigm 4, T. Lattimore and Szepesvári (2020) and Slivkins et al. (2019) offer the most recent comprehensive texts on bandit algorithms. These works cover a broad range of topics, with particular emphasis on algorithm design and regret analysis, including stochastic bandits, adversarial bandits, contextual bandits, etc. In the broader context of online policy learning, problems modeled as MDPs (paradigm 5) are typically studied through RL. Shakya et al. (2023) offers a comprehensive overview of RL fundamentals, while X. Wang et al. (2022) and Arulkumaran et al. (2017) focus on the integration of RL with deep learning. Gu, L. Yang, et al. (2024) surveys RL methods designed to address safety concerns in real-world applications, and Canese et al. (2021) and Gronauer and Diepold (2022) review multi-agent RL. For more complex settings such as POMDPs (paradigm 6), Xiang and Foo (2021) provides a detailed review of recent advances. However, these reviews generally overlook the connection between causality and policy learning.

Recently, the integration of causal knowledge into policy learning has garnered growing attention, leading to the emergence of the field of causal RL. For example, Schölkopf et al. (2021) briefly discusses the role and importance of causality in RL, with a main focus on causal representation learning. Kaddour et al. (2022) surveys causal machine learning and includes a brief chapter summarizing how RL can benefit from exploiting causal paradigms. Reviews by Y. Zeng et al. (2024) and Deng et al. (2023) are the most comprehensive, outlining how causal knowledge from causal discovery and causal inference can address key challenges faced by non-causal RL and systematically reviewing existing causal RL methods.

3 Preliminary

In this section, we provide a brief overview of the key concepts and assumptions that will be used throughout this paper. We begin with an introduction to the general principles of causal inference and policy learning. Next, we delve into the core concepts related to CSL, CEL, and CPL in Section 3.1-3.3, and conclude with the assumptions

outlined in Section 3.4. More specialized concepts specific to each task/paradigm are presented in their respective sections (Sections 5–8).

Definition 3.1. (Potential Outcome): For each individual, denote $A = a$ as the action or treatment assignment. We define $R(A = a)$ as the outcome or reward [1] if the individual receives action $A = a$. The potential outcomes framework, also known as the Neyman-Rubin Causal Model, is a foundational concept in causal inference.

Definition 3.2. (Do-Operator): Given any two variables X and Y in a causal system, the do-operator denotes an intervention on X, which is often written as $do(X = x)$. The conditional probability of Y given $do(X = x)$ is defined as $\mathbb{P}(Y|do(X = x))$.

Without additional assumptions about the causal structure involving (A, R), the probability $\mathbb{P}(R|do(A = a))$ generally differs from $\mathbb{P}(R|A = a)$. The discrepancy arises because $\mathbb{P}(R|do(A = a))$ represents the probability of R under an intervention where A is forcibly set to a, while all other potential causes of R, whether observed or not, are held fixed. Mathematically, if we denote Z as the set of all other variables that are causally upstream of R (excluding A), the intervention probability can be expressed as

$$\mathbb{P}(R|do(A = a)) = \sum_z \mathbb{P}(R|A = a, Z = z)\mathbb{P}(Z = z),$$

which captures how intervening on A with do-operator disrupts the natural causal mechanisms.

Definition 3.3. (Causal Identifiability) (Pearl 2009): A causal quantity Q (e.g., the average treatment effect, the causal graph, etc.) is said to be *identifiable* from a set of assumptions $\mathcal{AS}$ and observational data distribution $\mathcal{P}$, if for any two data-generating models $\mathcal{M}_1$ and $\mathcal{M}_2$ that satisfy the assumption $\mathcal{AS}$, the equality of their observed data distributions implies equality of the causal estimand:

$$\mathcal{P}_{\mathcal{M}_1} = \mathcal{P}_{\mathcal{M}_2} \quad \Rightarrow \quad Q_{\mathcal{M}_1} = Q_{\mathcal{M}_2}.$$

This definition states that a causal quantity is identifiable when it can be uniquely determined from the observed data distribution under the given assumptions. In other words, there exists a function g such that the causal quantity can be uniquely determined from the observed distribution, i.e., $Q = g(\mathcal{P})$. Within the framework of the Structural Causal Model (SCM), the assumption set $\mathcal{AS}$ typically encodes structural assumptions represented by a causal graph. In the potential outcomes framework, $\mathcal{AS}$ often corresponds to assumptions such as those listed in Section 3.4 (e.g., ignorability, consistency, positivity).

Definition 3.4. (Confounder): In causal structures, a variable C is considered a confounder between A and R if it acts as a common cause of both (A, R), i.e., $C \to A$ and $C \to R$.

Definition 3.5. (Mediator): A variable M is considered a mediator between A and R if it lies on the causal pathway from A to R, i.e. $A \to M \to R$.

Definition 3.6. (Decision Process): A decision process is a framework describing the evolution of states, actions, and rewards over time. In general settings, with the dataset being $\{s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_t, a_t, r_t, \ldots\}$, the probability of transitioning to a future state and receiving a reward can depend on the entire history of states and actions up to that point as $P(s_{t+1}, r_{t+1} \mid s_0, a_0, s_1, a_1, \ldots, s_t, a_t)$.

Definition 3.7. (Markov Decision Process (MDP)): An MDP is a special type of decision process in which the probability of transitioning to the next state and receiving a reward depends only on the current state and action, and not on any prior history. This "memoryless" (Markov) property simplifies decision making, as it allows the process to be fully characterized by the current state-action pair alone. Formally, $P(s_{t+1}, r_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = P(s_{t+1}, r_{t+1} \mid s_t, a_t)$.

[1] In the causal inference literature, the outcome is typically denoted by Y. To align with notation commonly used in reinforcement learning and to maintain consistency across both literatures, we use R to denote the outcome throughout this work.

3.1 Causal Graphical Model Under a Potential Outcome Framework

Consider a graph $\mathcal{G} = (X, D_X)$ with node set X and edge set D_X. There is at most one edge between any pair of nodes. If there exists an edge between X_i and X_j, then X_i and X_j are said to be *adjacent*. A node X_i is called a *parent* of X_j if there is a directed edge from X_i to X_j, i.e., X_i is a direct cause of X_j. Similarly, a node X_k is called an *ancestor* of X_j if there exists a directed path from X_k to X_j regulated by at least one additional node X_i for $i \neq k$ and $i \neq j$, i.e., X_k is an indirect cause of X_j. Let the set of all parents/ancestors of node X_j in $\mathcal{G}$ as $\text{PA}_{X_j}(\mathcal{G})$. A *path* from X_i to X_j in $\mathcal{G}$ is a sequence of distinct vertices, $\pi := \{a_0, a_1, \cdots, a_L\} \subset V$ such that $a_0 = X_i$, and $a_L = X_j$. A *directed path* from X_i to X_j is a path between X_i and X_j where all edges are directed towards X_j. A directed graph $\mathcal{G}$ that does not contain directed cycles is called a directed acyclic graph (DAG). A directed graph is acyclic if and only if it has a topological ordering.

The Structural Causal Model (SCM) characterizes causal relationships among $|X| = d$ nodes via a DAG $\mathcal{G}$ and noises $e_X = [e_{X_1}, \cdots, e_{X_d}]^\top$, such that $X_i := h_i\{\text{PA}_{X_i}(\mathcal{G}), e_{X_i}\}$ for some unknown h_i and $i = 1, \cdots, d$. Here, we allow the collection of nodes to play different causal roles within the causal graph. For example, let $A \in \mathbb{R}$ be an exposure/treatment; $M := (M_1, M_2, \cdots, M_p)^\top \in \mathbb{R}^p$ be mediators with dimension p in its support $M = \mathcal{M}_1 \times \cdots \times \mathcal{M}_p \subseteq \mathbb{R}^p$; and $R \in \mathbb{R}$ be the outcome of interest. Additionally, we also consider that there are $l - 1$ confounders $S := (S_1, \ldots, S_{l-1})^\top \in \mathbb{R}^{l-1}$ in its support $\mathcal{S} \subseteq \mathbb{R}^{l-1}$. We would just let $l = 1$ here to represent the absence of confounders, that is $\mathcal{S} = \emptyset$. Suppose there exists a DAG $\mathcal{G} = (X, D_X)$ that characterizes the causal relationships among $X = (S^\top, A, M^\top, R)^\top$, where the dimension of X is $d = l + p + 1$.

In such a scenario, the potential outcome framework (Rubin 1978) can be equivalently formulated using the 'do-operator' (Pearl 2000). For example, considering the reward R and action A, let $R(a) \equiv R(A = a)$ be the potential outcome that would be observed if the action A were set to a, following the notation of Rubin (1978). This term is stated to be equivalent to the value of R by imposing a 'do-operator' of $do(A = a)$ as in Pearl (2009):

$$R(a) \equiv R(A = a) \equiv R\{do(A = a)\},$$

where $do(A = a)$ is a mathematical operator that simulates a physical intervention that fixes A at value a while keeping the rest of the model unchanged. In graphical terms, this corresponds to removing all incoming edges to A and replacing A with the constant a in $\mathcal{G}$. Similarly, one can define the potential outcome, $R(X_i = x_i)$, by setting any variable X_i to a fixed value x_i while keeping the rest of the model unchanged. Suppose we observe data $X = (S^\top, A, M^\top, R)^\top$ for n subjects. The goal is to learn decision-oriented causal graphs $\mathcal{G}$ that capture the causal relationships among the variables in X based on the observed data.

3.2 Treatment Effect Estimation Under a Potential Outcome Framework

The fundamental challenge in causal inference is counterfactual estimation. Specifically, once a decision has been made and an action $A = a$ has been taken, we can only observe the outcome of that action. As a result, estimating the missing potential outcome $R(a')$ corresponding to the alternative action $A = a'$ becomes crucial. As a primary task in causal inference, treatment effect estimation can involve different concepts depending on the specific problem setting. Common causal estimands include the Average Treatment Effect (ATE), the Heterogeneous Treatment Effect (HTE), and the mediation effect.

Definition 3.8. (ATE): Under either the potential outcome framework or the do-operator system, the average treatment effect is defined as

$$\text{ATE} = \mathbb{E}[R(1) - R(0)] = \mathbb{E}[R|do(A = 1)] - \mathbb{E}[R|do(A = 0)].$$

For instance, when investigating the volume of fluids administered to patients with diabetes and its impact on their health status within 48 hours, the first question to address is, "Is this intravenous (IV) fluid generally

effective in reducing the mortality rate?" This question pertains to estimating the treatment effect on the overall patient population, which is quantified by the ATE as defined above.

Definition 3.9. (HTE): To account for the heterogeneous effects across different individuals or contextual groups, the HTE is defined as

$$\tau(s) = \mathbb{E}[R(1) - R(0)|S = s] = \mathbb{E}[R|do(A = 1), S = s] - \mathbb{E}[R|do(A = 0), S = s].$$

Unlike the ATE, which focuses on the overall effect across the population, the HTE further explores the variation in treatment effects across different subgroups or individuals. In the diabetes example, the HTE aims to understand whether IV fluid administration leads to different levels of causal effects for patients with varying characteristics, such as age, gender, or prescription history, as captured by the state variable S.

Definition 3.10. (Mediation Effect): When mediators are involved, the total effect (TE) can be decomposed into the natural direct effect (DE) and the natural indirect effect (IE), where

$$\begin{aligned}
\text{TE} &= \mathbb{E}[R|do(A = a_1)] - \mathbb{E}[R|do(A = a_0)] \\
\text{DE} &= \mathbb{E}[R|do(A = a_1, M = m^{(a_0)})] - \mathbb{E}[R|do(A = a_0), M = m^{(a_0)}] \\
\text{IE} &= \mathbb{E}[R|do(A = a_1, M = m^{(a_1)})] - \mathbb{E}[R|do(A = a_0, M = m^{(a_1)})]
\end{aligned}$$

with TE = DE + IE.

In the context of diabetes, the Sepsis-related Organ Failure Assessment (SOFA) score can be considered a mediator influenced by the administration of IV fluid. This score, in turn, affects the mortality rate within the next 48 hours. The mediation effect allows decomposing the total effect of IV fluid on mortality into two components: the direct effect, which directly measures how IV fluid impacts mortality, and the indirect effect, which operates through the SOFA score to influence mortality.

3.3 Causal Policy Learning Under a Potential Outcome Framework

Definition 3.11. (Policy): A policy π is the agent's strategy, defined as a function that maps relevant information (context/state in Paradigm 1-2 or 4-5; all historical information in non-Markovian decision processes) to an action (for deterministic policies) or to a probability distribution over actions (for stochastic policies).

We commonly use the value functions to evaluate the goodness of a policy. Here, we illustrate this under a discounted infinite-horizon setting.

Definition 3.12. (V-function): The state value function (V-function) under policy π is defined as:

$$V^{\pi}(s) = \sum_{t \geq 0} \gamma^t \mathbb{E}_{\pi}\{R_t|S_0 = s\},$$

where $0 < \gamma < 1$ is a discount factor reflecting the trade-off between immediate and future rewards. The value function measures the discounted cumulative reward the agent would receive if it followed policy π.

Definition 3.13. (Q-function): The state-action value function (Q-function) under policy π is defined as:

$$Q^{\pi}(a, s) = \sum_{t \geq 0} \gamma^t \mathbb{E}_{\pi}\{R_t|S_0 = s, A_0 = a\}.$$

Definition 3.14. (Optimal Policy): The optimal policy, π^*, is defined as

$$\pi^* = \arg\max_{\pi} V^{\pi}(s), \forall s \in \mathcal{S}.$$

Definition 3.15. (Bellman Optimality Equations): The Q-learning-type policy learning is commonly based on the Bellman optimality equation, which characterizes the optimal policy π^* and underlies many policy optimization methods. Specifically, Q^* is the unique solution of

$$Q(a,s) = \mathbb{E}\Big(R_t + \gamma \arg\max_{a'} Q(a, S_{t+1}) | A_t = a, S_t = s\Big). \tag{1}$$

The concepts above can also be extended to the non-Markovian settings, with the state variable replaced by its full history.

3.4 Three Key Causal Identifiability Assumptions

To address the problem of counterfactual estimation, causal inference typically relies on three key assumptions. While recent research has focused on relaxing these assumptions, we first detail them here and discuss scenarios where these assumptions may be violated in Section 9.

Assumption 3.1. *Stable Unit Treatment Value Assumption (SUTVA) states that*

$$R_i = \sum_{A=a} R_i(a) 1\{A_i = a\}, i \in \{1, \cdots, n\}, \tag{2}$$

which can be divided into two key sub-assumptions: (i) No interference between units, meaning that the potential outcomes for one unit are unaffected by the actions assigned to other units; and (ii) Consistency of treatment, meaning that there are no different versions of the same action that could lead to different potential outcomes.

Assumption 3.2. *No Unmeasured Confounders (NUC) assumes that*

$$R(a) \perp\!\!\!\perp A | S, \quad \forall a \in \mathcal{A},$$

which quantifies the conditional independence of potential outcomes from the action being taken.

For example, when investigating whether regular exercise reduces the risk of heart disease, genetic factors might influence both a person's likelihood to exercise and their risk of heart disease. In this case, genetic predisposition acts as an unmeasured confounder that violates the NUC assumption by affecting both the treatment (exercise) and the outcome (heart disease risk).

Assumption 3.3. *The positivity (or overlap) assumption states that*

$$0 < c_0 < P(A = a|S) < c_1 < 1, \quad \forall a \in \mathcal{A},$$

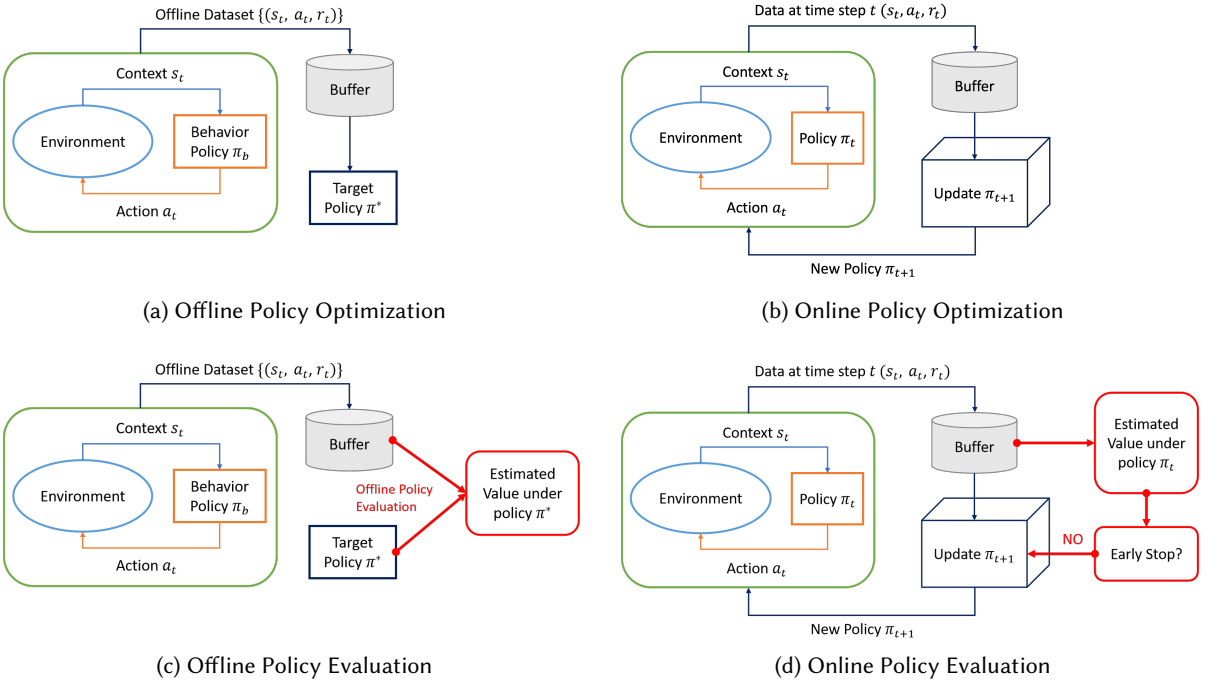
which assumes that every unit in the study population has a nonzero probability of receiving each possible treatment or intervention.

4 Core Concepts for CDM

In this section, we introduce the core concepts of CDM, including the three fundamental decision-making tasks and the six paradigms commonly assumed for the data-generating process underlying decision-making problems.

4.1 Three Tasks

Causal Structure Learning. In recent years, *causal discovery*, also known as CSL (e.g., Pearl 2000; Peters, Janzing, et al. 2017), has gained great attention for disentangling complex causal relationships in many areas (e.g., Cai et al. 2020; Chakrabortty et al. 2018; Nandy, Maathuis, et al. 2017). Building upon causal graphical models, (see Section 3.1 and Pearl 2009, for a comprehensive review), several CSL methods have been proposed (e.g., Bühlmann et al. 2014; Kalisch and Bühlmann 2007; Shimizu et al. 2006; Spirtes, C. Glymour, et al. 2000; Yu et al. 2019; X. Zheng, Aragam, et al. 2018; S. Zhu et al. 2020) to estimate the causal graphs from observed data,



(a) Offline Policy Optimization

(b) Online Policy Optimization

(c) Offline Policy Evaluation

(d) Online Policy Evaluation

Fig. 3. Causal Policy Learning

which is a crucial step in understanding the underlying mechanisms that govern changes in a system. It involves identifying the causal relationships between variables (see an illustration in the first panel of Figure 1), which is fundamental for any subsequent analysis aiming to understand causal effects and make causal decisions. Most existing methodologies for average/heterogeneous treatment effects, as well as personalized decision making, rely on a known causal structure, which provides the convenience of identifying the right variables to control (e.g., confounders), to intervene (e.g., treatments), and to optimize (e.g., rewards). However, such convenience is absent in many emerging real applications where the causal structure is unknown. Causal discovery has therefore attracted increasing attention recently for inferring causal structures from data and disentangling the complex relationships among variables. In Section 5, we present state-of-the-art techniques for learning the skeleton of unknown causal relationships among input variables with embedded treatments.

Causal Effect Learning. CEL is a process of determining cause-and-effect relationships between variables (Pearl 2009, 2003). Given a causal structure, either pre-assumed based on domain knowledge or derived using CSL methods, the goal of CEL is to identify, estimate, and infer the causal effect of interest. CEL primarily focuses on estimating several key effects, including the Average Treatment Effect (ATE) (Hirano et al. 2003; G. W. Imbens 2004), which measures the overall impact of a treatment across the entire population; the Heterogeneous Treatment Effect (HTE) (Curth and Schaar 2021; Wager and Athey 2018), which captures how the treatment's effect varies across different subgroups or individuals; and the mediation effect, which decomposes the causal effect by accounting for intermediate variables or mediators (T. J. VanderWeele 2016) that influence the relationship between the treatment and the outcome. In Section 6, we provide a comprehensive review of the literature on CEL, covering diverse paradigms and data structures.

Causal Policy Learning. With the causal structure and effects between variables in mind, the ultimate goal is typically to evaluate and optimize our decision making. When decision making is purely based on a fixed historical dataset (i.e., it does not involve continuous data collection), we refer to such a setting as *offline* or *off-policy* (see Figures 3a and 3c); whereas the decision-making process involves continuous data collection and real-time policy updates based on incoming outcomes is referred to as *online* (see Figures 3b and 3d). Regardless of the data collection method, two fundamental tasks in CPL are *policy evaluation* and *policy optimization*. Policy evaluation (Y. Shen et al. 2024; Uehara, Shi, et al. 2022; Voloshin et al. 2019) involves estimating the value of a given *target policy* with respect to a specific state distribution (see Figure 3c) or assessing the value of the estimated optimal policy in online learning (see Figure 3d). Policy optimization (Bouneffouf et al. 2020; Ladosz et al. 2022; Levine, Kumar, et al. 2020; Moerland et al. 2023; Prudencio et al. 2023; Shakya et al. 2023; Silva et al. 2022; X. Wang et al. 2022) focuses on determining the optimal policy that maximizes its value under certain problem-specific requirements (see Figure 3a) or identifying the optimal actions that maximize the cumulative rewards during online interactions (see Figure 3b). In Section 7, we review the literature on offline CPL, while online CPL is discussed in Section 8.

4.2 Six Paradigms

Regardless of the specific CDM task, decision-making problems in the literature can be categorized into six paradigms, each capturing common data dependencies typically assumed in practice, as illustrated in Figure 2 and detailed below.

Paradigm 1: Fixed Policy with Independent States. As illustrated in Figure 2, observations in Paradigm 1 are i.i.d. samples. Each observation consists of three components: S_i is the contextual information (if available), A_i is the action taken, and R_i is the reward received. When contextual information is present, it would influence the choice of action, while both the contextual information and the action jointly determine the final reward. A classical class of problems that are widely studied in this context is the single-stage DTR (Tsiatis et al. 2019). Literature on CSL within this paradigm is discussed in Section 5.3, while studies on ATE, HTE, and mediation effect analysis are reviewed in Section 6.2. Additionally, this paradigm serves as the main setting in Section 7 to illustrate offline policy learning methods for evaluating or learning (personalized) policies that aim to maximize the immediate rewards.

Paradigm 2: Fixed Policy with Markovian State Transition. Paradigm 2 is widely recognized as the Markov Decision Process (MDP), characterized by Markovian state transitions. In particular, while A_t is only affected by S_t, both R_t and S_{t+1} are affected by the pair (S_t, A_t). Given S_t and A_t, a standard assumption in MDP problems is that R_t and S_{t+1} are conditionally independent of all previous observations. In Section 7, we also extend the policy evaluation and optimization techniques developed in Paradigm 1 to this setting, where the policy aims to maximize the long-term rewards.

Paradigm 3: Fixed Policy with Non-Markovian State Transition. Paradigm 3, commonly assumed in multi-stage DTR problems (Tsiatis et al. 2019) and offline non-Markovian RL problems, considers all possible causal relationships under a history-independent policy. CPL-related studies within this paradigm are briefly reviewed in Section 7. In the context of CEL within Paradigm 3, we primarily focus on a specific panel data setting and discuss effect estimation with respect to evolving time, as detailed in Section 6.4.

Paradigm 4: Adaptive Policy with Independent States. Paradigm 4 is extensively studied in the online decision-making literature as the bandit problem, where the treatment policy is time-adaptive. In this paradigm, the history H_{t-1}, which includes all prior observations up to time $t-1$, is used to update the action policy at time t, thereby influencing the decision-making for action A_t. While S_t is sampled i.i.d. from the corresponding distribution, the reward R_t is influenced by both A_t and S_t. The new observation (S_t, A_t, R_t), combined with all previous observations, forms the updated history H_{t+1}, which then affects the next action A_{t+1}. A common

Table 1. Summary of Causal Structure Learning Literature

Application	Method Type	Model	Papers
Decision-Oriented CSL under Paradigm 1	Testing-based	PC algorithm extended with treatment-outcome ordering	Chakrabortty et al. (2018), Maathuis et al. (2009), and Nandy, Maathuis, et al. (2017)
	Score-based	SEM-based optimization with acyclicity constraint and identification penalty	Cai et al. (2020) and Watson et al. (2023)
Decision-Oriented Mediation Analysis	Testing-based	PC with linear SEMs	Chakrabortty et al. (2018) and T. VanderWeele and Vansteelandt (2014)
	Score-based	DAG-based causal mediation with constrained estimation	Cai et al. (2020) and Watson et al. (2023)
Time-Series CSL (Paradigm 2+)	Testing-based	Lagged CI test with autocorrelation correction	Runge (2018) and Runge et al. (2019)
	Functional-based	Attention-based temporal graph discovery and neural DAG discovery	Nauta et al. (2019) and Tank et al. (2021)

variation of this structure occurs when the contextual information S_t is absent. Relevant literature on policy optimization under Paradigm 4 is reviewed in Section 8.1.1, while related policy evaluation approaches are discussed in Section 8.2.

Paradigm 5: Adaptive Policy with Markovian State Transition. Building on Paradigm 4, Paradigm 5 introduces the MDP framework, with an adaptive policy and Markovian state transitions governing the data generation process. Specifically, S_t follows a Markovian state transition that depends solely on the most recent state and action and A_t is determined by the entire observation history H_{t-1} through a dynamically updated action policy. This setup corresponds to the typical online RL setup, which is reviewed in Section 8.1.2.

Paradigm 6: Adaptive Policy with Non-Markovian State Transition. Paradigm 6 extends Paradigm 5 by relaxing the Markovian assumption, allowing for non-Markovian state transitions. This paradigm encompasses problems such as the Partially Observable Markov Decision Process (POMDP), with relevant approaches briefly reviewed in Section 8.1.2.

5 Causal Structure Learning

Most existing methodologies for average/heterogeneous treatment effects and personalized decision making rely on a known causal structure. This enables us to identify the right variables to control (e.g., confounders), to intervene (e.g., treatments), and to optimize (e.g., rewards). However, such convenience is often violated in many emerging real-world applications with unknown causal structures. Causal discovery has therefore attracted more and more attention recently, as it allows inferring causal structure from data and disentangling complex relationships among variables. In this section, we present state-of-the-art techniques for learning the skeleton of unknown causal relationships among input variables with embedded treatments. We first detail why CSL is needed for causal decision making, then introduce common causal graphical models in Paradigm 1 and present existing representative classes of causal discovery methods, followed by discussions of their extensions to Paradigms 2 and 3.

5.1 Why CSL is Needed for Causal Decision Making

CSL is a crucial step in understanding the underlying mechanisms that govern changes in a system. It involves identifying the causal relationships between variables, which is fundamental for any subsequent analysis aiming to understand causal effects and support causal decision making. The important reasons why CSL precedes CEL and CPL can be summarized as follows.

First, CSL is the *first* step of causal decision making. More specifically, CSL is essential for *designing effective interventions and policies* by identifying the exposure or treatment in the causal graph (see e.g., Bühlmann et al. 2014; Chickering 2002; N. Harris and Drton 2013; Kalisch and Bühlmann 2007; Ramsey et al. 2017; Shimizu et al. 2006; Spirtes, C. Glymour, et al. 2000; J. Zhang and Bareinboim 2018b). In fields such as epidemiology (M. A. Hernán 2004), medicine (M. Á. Hernán et al. 2000), and economics (Panizza and Presbitero 2014), the underlying causal mechanisms among variables of interest are typically unknown. CSL allows intervention evaluators or policymakers to understand the potential ramifications of their actions by revealing how different factors interact causally.

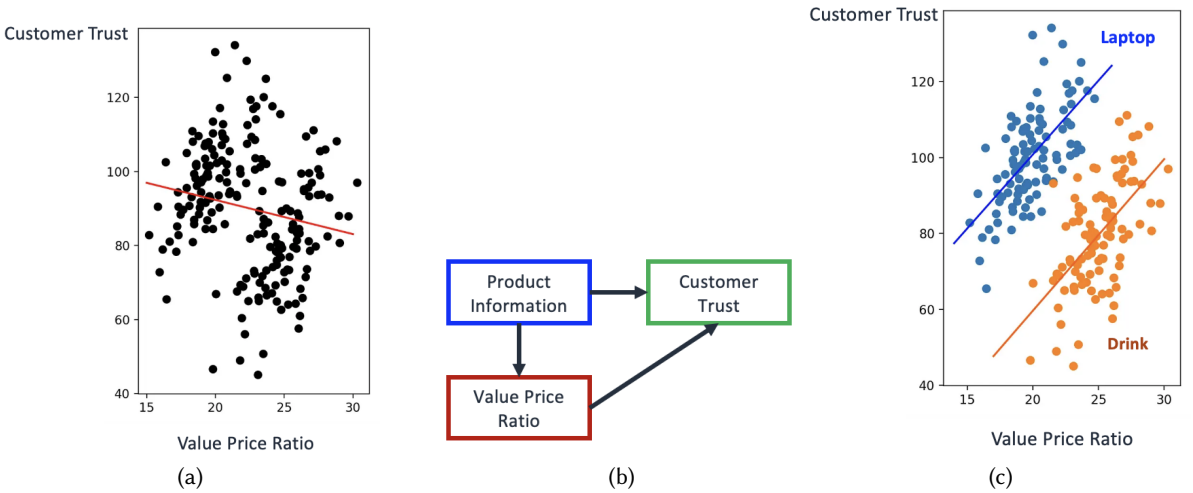
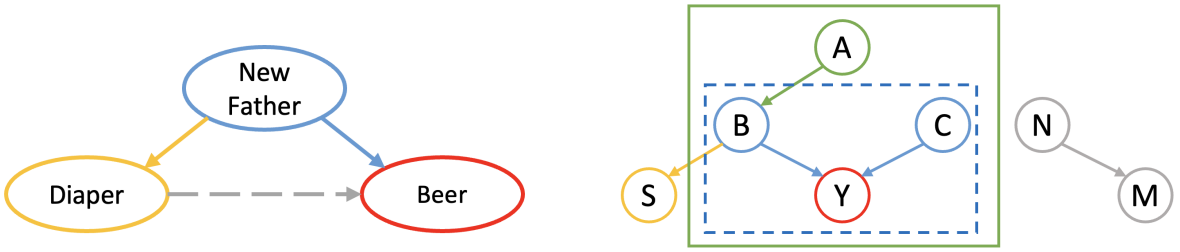


Fig. 4. An illustration of Simpson's Paradox. **(a)**: A simplistic regression that ignores confounders. **(b)**: A complete causal graph that accounts for potential confounders. **(c)**: A corrected regression that considers all confounders.

Second, understanding the causal pathway is essential when estimating the impact of changes in one variable on another to inform decision making. Interventions based on *incomplete causal knowledge risk yielding biased outcomes*, as depicted in Figure 4. This figure illustrates the complexities of discerning the relationship between the value-price ratios and customer trust. A simplistic regression that ignores confounders suggests a counterintuitive negative correlation: higher value-price ratios correspond to lower customer trust, as shown in Figure 4a. However, this analysis is flawed due to omitted variable bias. Figure 4b introduces a complete causal graph that accounts for potential confounders, offering a more accurate representation of the relationship. When product information is incorporated, the apparent contradiction resolves—higher value-price ratios actually correlate with increased trust in both laptop and drink product categories, as illustrated in Figure 4c with the corrected regression. This scenario exemplifies Simpson's Paradox (Blyth 1972), where aggregated data can mask or reverse trends present within stratified groups.



(a) (b)

Fig. 5. **(a)**: Illustration of the causal relationship between the customer being a new father or not, beer purchasing, and diaper purchasing, where solid lines represent the true model, and the dashed line corresponds to the spurious correlation between beer purchasing and diaper purchasing. **(b)**: Relationship between various causal structures. Nodes A, B, and C belong to the necessary and sufficient causal graph for the target outcome Y and are depicted inside the green solid square. Among them, nodes B and C are direct parents of Y, enclosed by the blue dotted square. Node S is the spurious variable to Y, while nodes N and M are unrelated to the target.

Third, with CSL, we can *avoid spurious relationships*. Building upon the causal graphical model (see e.g., Pearl 2009), many CSL algorithms have been developed (see e.g., Bühlmann et al. 2014; Cai et al. 2020; Chickering 2002; Kalisch and Bühlmann 2007; Ramsey et al. 2017; Shimizu et al. 2006; Spirtes, C. Glymour, et al. 2000; Yu et al. 2019; X. Zheng, Aragam, et al. 2018; S. Zhu et al. 2020), but they rely on the assumption of causal sufficiency (the absence of unmeasured confounders). In real-world applications, to satisfy such an assumption, we strive to learn large-scale causal graphs (see e.g., Chakrabortty et al. 2018; Nandy, Maathuis, et al. 2017; Niu et al. 2021; Tang et al. 2020), in the hope of sufficiently describing how an outcome of interest depends on its relevant variables. In addition to sufficiency, it is also crucial to account for the concept of necessity by excluding redundant variables that explain the outcome of interest. Failure to do so can result in the inclusion of spurious variables in the learned causal graphs, which are highly correlated with, but have no causal impact on, the outcome. These variables can impede causal estimation with limited data and lead to falsely discovered spurious relationships, resulting in poor generalization performance for downstream prediction (Schölkopf et al. 2021). For instance, it might be observed that men aged 30 to 40 who buy diapers are also likely to purchase beer. However, beer purchase is a spurious feature for diaper purchase: their correlation is not necessarily causal, as both purchases might be confounded by a shared cause, such as new fathers who buy diapers for childcare while also buying beer to alleviate stress. Therefore, merely increasing the availability of diapers or beer will not causally enhance the demand for the other (see also Figure 5a).

Fourth, with CSL, we aim to *simplify complex models* by identifying the most relevant causal relationships for decision making. This simplification can make models more understandable, efficient, and less prone to overfitting. The number of variables causally relevant to the outcome of interest is often considerably smaller than the number of variables included when estimating a causal graph (see Figure 5b). For example, while an individual's genome may encompass 4 to 5 million single nucleotide polymorphisms (SNPs), only a limited number of non-spurious genes or proteins are found to systematically regulate the expression of the phenotype of interest (e.g., Chakrabortty et al. 2018). Similarly, in natural language processing tasks, excluding spurious embeddings such as writing style and dialect can enhance model accuracy and downstream prediction performance (e.g., Feder et al. 2022).

5.2 Overview of Decision-Oriented Causal Discovery

Under a general treatment-embedded causal graph, the treatment or exposure may have a direct effect on the outcome and also an indirect effect regulated by a set of mediators (or intermediate variables), confounded by some baseline covariates. In the era of the causal revolution, identifying the causal effect of the exposure on the outcome of interest is an important problem in many areas (see e.g., Cai et al. 2020; Chakrabortty et al. 2018; Watson et al. 2023). An analysis of causal effects that interprets the causal mechanism contributed through mediators is hence challenging but on demand, and naturally bridges the gap between CSL and CEL, with the learned results further serving as the middle step for CPL.

Existing statistical and machine learning tools for learning the causal graphs with multiple mediators (see e.g., Cai et al. 2020; Chakrabortty et al. 2018; Shi and L. Li 2021) comprise the following three principal steps. Initially, CSL methodologies (see e.g., Chickering 2002; C. Li et al. 2023, 2019; Nandy, A. Hauser, et al. 2018; Spirtes, C. Glymour, et al. 2000; Yuan et al. 2019) are applied to estimate the causal graph, often presented by a DAG, using observational data. With preliminaries of the causal graphical model (Pearl 2009; Peters and Bühlmann 2014) introduced in Section 5.3.1, we present decision-oriented CSL methods by extending three representative classes of causal discovery methods in Section 5.3.2. In the absence of additional assumptions (Neal 2020; Shimizu et al. 2006), the graph is only identified up to a Markov equivalence class (MEC), and a completed partially directed acyclic graph (CPDAG) in such a class is often used to represent the graph structure (see details in Section 5.3.3). Following the advances in causal mediation analysis (Cai et al. 2020) with complex graph structure (see details in Section 5.3.4), the subsequent step is the estimation of the causal effects of mediators based on the DAG or CPDAG obtained from the initial phase. For this task, a variety of estimation techniques have been proposed, including the application of ordinary least squares (OLS) estimators (Chakrabortty et al. 2018; S.-H. Lin and T. VanderWeele 2017; T. J. VanderWeele and W. R. Robinson 2014), parametric models (L. Chen et al. 2024; T. VanderWeele and Vansteelandt 2014; T. J. VanderWeele, Jackson, et al. 2016), and nonparametric methods (An and T. J. VanderWeele 2022; Brand et al. 2023).

5.3 Decision-Oriented CSL Under Paradigm 1

This section presents state-of-the-art techniques for learning the skeleton of unknown causal relationships among input variables with the presence of treatments or decision variables under Paradigm 1.

5.3.1 Preliminaries in Decision-Oriented Causal Discovery Under Paradigm 1. As commonly imposed in works of CSL (e.g., Peters, Mooij, et al. 2014; Spirtes, C. N. Glymour, et al. 2000), we assume the causal Markov and faithfulness assumptions. To detail these assumptions, we first introduce the concept of the D-separation.

Definition 5.1 (D-separation)**.** Nodes, X and Y, are d-separated by a set of nodes, Z, if and only if for every path, π, there exists a node, $m \in Z$, that extends π ($i \to m \to j$) or forks π ($i \leftarrow m \to j$) and for any node, c, along π that is a so-called collider ($i \to m \leftarrow j$), c and all descendants of c are not in Z (Pearl 2009).

Given that Z d-separates X and Y, and X precedes Y causally, the implication of d-separation is that $X \perp Y|Z$.

Assumption 5.1 (Causal Markov assumption)**.** *For a given causal graph, $\mathcal{G} = (\mathbf{Z}, \mathbf{E})$, the set of independences among the nodes, $\mathbf{Z}$, contains the set of independences implied by d-separation.*

Assumption 5.2 (Faithfulness assumption)**.** *For a given causal graph, $\mathcal{G} = (\mathbf{Z}, \mathbf{E})$, the set of independences among the nodes, $\mathbf{Z}$, is **exactly** described by the set of independences implied by applying d-separation to $\mathcal{G}$.*

Note that the assumptions made in this review paper are commonly imposed in the literature of causal inference. Please refer to Athey and G. W. Imbens (2015), Künzel et al. (2019), Nandy, Maathuis, et al. (2017), Nie and Wager (2021), Pearl (2009, 2000), and Wager and Athey (2018) for discussions of these assumptions and their impact. There are a few future extensions to relax or diagnose these assumptions. For instance, a full sensitivity analysis

of the assumptions would be useful to the field when it is hard to include all variables causally related to any variable in the data in practice. In addition, utilizing the instrumental variables in the context of causal graphs with multiple mediators may be beneficial in addressing unmeasured confounders, as specified below.

Assumption 5.3 (Causal Sufficiency assumption)**.** *The causal graph $\mathcal{G}$ satisfies Causal Sufficiency (Hasan et al. 2023). The random vector* $\mathbf{X}$ *satisfies the structure assumption: (i) No potential mediator is a direct cause of confounders* $\mathbf{S}$*; (ii) The outcome R has no descendant; (iii) The only parents of treatment A are confounders.*

In many instances, the accessible data offer an incomplete view of the inherent causal structure. To address this gap, the *Causal Markov Condition*, *Causal Faithfulness Condition*, and *Causal Sufficiency* in the above assumptions provide soundness and completeness for causal discovery in i.i.d. data contexts (Assaad et al. 2022; Hasan et al. 2023; S. Lee and Honavar 2020). The rigorous definitions of these conditions and related details can be found in Section 2.4 of Hasan et al. (2023). Furthermore, the structural assumptions for decision-oriented CSL aim at ensuring the identifiability of the causal model, which are similar to the Consistency Assumption and the Sequential Ignorability Assumption in Tchetgen and Shpitser (2012), and the structure assumptions in Section 2.4 of Chakrabortty et al. (2018). We next introduce some commonly considered causal graphical models under Paradigm 1, as follows.

Linear Structural Equation Model. Let $\boldsymbol{B} = \{b_{i,j}\}_{1\le i\le d, 1\le j\le d}$ be a $d \times d$ matrix, where $b_{i,j}$ is the weight of the edge $X_i \longrightarrow X_j \in \mathbf{E}$, and $b_{i,j} = 0$ otherwise. Then, we say that $\mathcal{G} = (\boldsymbol{X}, \mathbf{E})$ is a weighted DAG with the node set $\boldsymbol{X}$ and the weighted adjacency matrix $\boldsymbol{B}$ (the edge set $\mathbf{E}$ is nested in $\boldsymbol{B}$). The linear structural equation model (LSEM) (Sobel 1987) such that $\boldsymbol{X} = (\boldsymbol{S}^\top, A, \boldsymbol{M}^\top, R)^\top$ characterized by the pair $(\mathcal{G}, \epsilon)$ is generated by

$$\boldsymbol{X} = \boldsymbol{B}^\top \boldsymbol{X} + \epsilon, \tag{3}$$

where ϵ is a random vector of jointly independent error variables. We next explicitly characterize the weighted adjacency matrix $\boldsymbol{B}$ that satisfies Model (3) based on causal knowledge among $\boldsymbol{S}, A, \boldsymbol{M}$, and R, in the decision-oriented CSL. Specifically, the following matrix $\boldsymbol{B}^\top$ consists of unknown parameters whose sparsity is due to prior causal information:

$$\boldsymbol{B}^\top = \begin{bmatrix} \mathbf{0}_{l\times l} & \mathbf{0}_{l\times 1} & \mathbf{0}_{l\times p} & \mathbf{0}_{l\times 1} \\ \boldsymbol{\delta}_S & 0 & \mathbf{0}_{1\times p} & 0 \\ \boldsymbol{B}_S{}^\top & \boldsymbol{\beta}_A & \boldsymbol{B}_{\boldsymbol{M}}{}^\top & \mathbf{0}_{p\times 1} \\ \boldsymbol{\gamma}_S & \gamma_A & \boldsymbol{\gamma}_{\boldsymbol{M}} & 0 \end{bmatrix},$$

where $\mathbf{0}_{a\times b}$ is a $a \times b$ zero matrix/vector, and the parameters $\boldsymbol{\delta}_S, \boldsymbol{B}_S{}^\top$, and $\boldsymbol{\gamma}_S$ represent the influence of $\boldsymbol{S}$, on the treatment A, the mediators $\boldsymbol{M}$, and the outcome R, respectively. Likewise, $\boldsymbol{\beta}_A$ and γ_A represent the influence of A on $\boldsymbol{M}$ and R, respectively, and $\boldsymbol{\gamma}_{\boldsymbol{M}}$ represent the influence of $\boldsymbol{M}$ on R. $\boldsymbol{B}_{\boldsymbol{M}}{}^\top$ represents the influence of the mediators on other mediators. If $\boldsymbol{B}_{\boldsymbol{M}}{}^\top = \mathbf{0}_{p\times p}$ for the p-dimensional mediators, then we say that mediators are *parallel*, otherwise they are *sequentially ordered*. The extension to the LSEM with the interaction between the possible moderators and the treatment can be found in Watson et al. (2023).

Additive Noise Model. Suppose there exists a weighted DAG $\mathcal{G} = (\boldsymbol{X}, \mathbf{E})$ that characterizes the causal relationship among $|\boldsymbol{X}| = d$ nodes in $\boldsymbol{X} = (\boldsymbol{S}^\top, A, \boldsymbol{M}^\top, R)^\top$. Each variable X_i is associated with a node i in the DAG $\mathcal{G}$, and the observed value of X_i is obtained as a function of its parents in the graph plus an independent additive noise n_i, as the additive noise model (Bühlmann et al. 2014), i.e.,

$$X_i := f_i\{PA_{X_i}(\mathcal{G})\} + n_i, i = 1, 2, ..., d, \tag{4}$$

where $PA_{X_i}(\mathcal{G})$ denotes the set of parent variables of X_i so that there is an edge from $X_j \in PA_{X_i}(\mathcal{G})$ to X_i in the graph, and the noises n_i are assumed to be jointly independent. Here, Model (3) is a special case of Model (4).

Generalized LSEM. To handle complex relationships, a generalized version of LSEM has been studied (Yu et al. 2019) as

$$f_2(\boldsymbol{X}) = B^\top f_2(\boldsymbol{X}) + f_1(\epsilon), \tag{5}$$

where the parameterized functions f_1 and f_2 effectively perform (possibly nonlinear) transforms on ϵ and $\boldsymbol{X}$, respectively. Here, Model (3) is also a special case of Model (5).

5.3.2 Decision-Oriented Causal Discovery Methods Under Paradigm 1. In this section, we mainly focused on the decision-oriented CSL methods under Paradigm 1. Plentiful CSL methods have been proposed, with the large literature categorized into three types, including testing-based learners (Kalisch and Bühlmann 2007; Spirtes, C. Glymour, et al. 2000), functional-based learners (Bühlmann et al. 2014; Shimizu et al. 2006), and score-based learners (Ramsey et al. 2017; Yu et al. 2019; X. Zheng, Aragam, et al. 2018; S. Zhu et al. 2020). However, all these methods treat nodes in the graph as generic variables without any causal meaning. In the following, we review a few recent works that learn causal graphs with decision variables oriented.

To start with, we briefly introduce the Peter-Clark (PC) algorithm (Spirtes, C. N. Glymour, et al. 2000), named by the first two authors, Pater and Clark, as one of the oldest testing-based (or constraint-based) algorithms for causal discovery, and the existing decision-oriented CSL methods based on the Peter-Clark (PC) algorithm. To learn the underlying causal structure, the PC algorithm depends largely on conditional independence (CI) tests. If two variables are statistically independent or conditionally independent, there is no causal link between them. Maathuis et al. (2009) started using an unknown DAG without hidden variables to estimate the causal effects from the high-dimensional observational data based on the PC algorithm. Later, Nandy, Maathuis, et al. (2017) extended the work of Maathuis et al. (2009) with a linear structural equation model. More recently, following these works, Chakrabortty et al. (2018) first introduced the treatment or decision variable into the linear structural equation model and further defined the individual mediation effect. To identify such a causal graph, Chakrabortty et al. (2018) fixed the first variable as the treatment or decision and the last variable as the outcome of interest, and then applied the PC algorithm to the rest of the model, i.e., the multiple mediators that influence the outcome but are controlled by the treatment. More specifically, their algorithm finds and orients the v-structures or colliders (i.e., $X \rightarrow Y \leftarrow Z$) based on the d-separation set of node pairs (see Definition 5.1). All of these models rely on the PC algorithm to search for the Markov equivalence class of the partial DAG, and usually require strong sparsity and normality assumptions due to computational limits.

Next, we focus on another type of causal discovery approaches, the score-based methods, including greedy equivalence search (Chickering 2002; B. Huang et al. 2018; Ramsey et al. 2017) and acyclicity optimization methods (Cai et al. 2020; Lachapelle et al. 2020; Vowels et al. 2021; Yu et al. 2019; X. Zheng, Aragam, et al. 2018; X. Zheng, Dan, et al. 2020; S. Zhu et al. 2020). In the following, we detail a score-based learner, NOTEARS (X. Zheng, Aragam, et al. 2018), as an example and then extend to recent decision-oriented CSL methods. X. Zheng, Aragam, et al. (2018) constructed an optimization with an acyclicity constraint under the LSEM, i.e. the NOTEARS. A follow-up work using a variational autoencoder parameterized by a graph neural network that generalizes LSEM was proposed in Yu et al. (2019) with a more computationally friendly constraint, namely DAG-GNN. See also S. Zhu et al. (2020) and Lachapelle et al. (2020) for other cutting-edge score-based structural learning methods. However, these methods cannot be directly applied to decision-oriented causal graphs. To address this challenge, Cai et al. (2020) proposed a new constrained structural learning approach by incorporating the background knowledge (the temporal causal relationships among variables) into the score-based algorithms. They formulated such prior information as an identification constraint and added it as a penalty term in the objective function for the causal discovery. In the following, we typically detail the NOTEARS for illustration, which can be easily extended to other score-based algorithms. Specifically, as an example, we can write the linear structural model in

Equation (3) under the causal sufficiency assumption without states as

$$\begin{bmatrix} A \\ \boldsymbol{M} \\ R \end{bmatrix} = \boldsymbol{B}^\top \begin{bmatrix} A \\ \boldsymbol{M} \\ R \end{bmatrix} + \epsilon = \begin{bmatrix} 0 & \mathbf{0}_{p\times 1} & 0 \\ \boldsymbol{\alpha} & B_M^\top & 0 \\ \gamma & \boldsymbol{\beta}^\top & 0 \end{bmatrix} \begin{bmatrix} A \\ \boldsymbol{M} \\ R \end{bmatrix} + \begin{bmatrix} \epsilon_A \\ \epsilon_{M_p} \\ \epsilon_R \end{bmatrix}, \tag{6}$$

where γ is a scalar, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\mathbf{0}_{p\times 1}$ are $p \times 1$ vectors, B_M is a $p \times p$ matrix, and $\epsilon \equiv [\epsilon_A, \epsilon_M^\top, \epsilon_R]^\top$. Here, γ presents the weight of the edge $A \to R$, the i-th element of $\boldsymbol{\alpha}$ corresponds to the weight of the edge $A \to M_i$, and the i-th element of $\boldsymbol{\beta}$ is the weight of the edge $M_i \to R$. Note that by the causal sufficiency assumption, we have that exposure A has no parents and the outcome R has no descendants, so equivalently, the first row and the last column of $\boldsymbol{B}^\top$ are all zeros (i.e., the first column and the last row of $\boldsymbol{B}$ are all zeros). To estimate the weighted adjacency matrix B, the score-based learners formulate the acyclicity constraint (Yu et al. 2019; X. Zheng, Aragam, et al. 2018) as $h_1(\boldsymbol{B}) \equiv \operatorname{tr}\left[(I_{d+1} + t\boldsymbol{B} \circ \boldsymbol{B})^{d+1}\right] - (d+1) = 0$, where I_{d+1} is a $d+1$-dimensional identity matrix, $\operatorname{tr}(\cdot)$ is the trace of a matrix, and t is a hyperparameter that depends on the estimated largest eigenvalue of $\boldsymbol{B}$. The task of learning DAG is transformed into a constrained optimization problem with the loss by the augmented Lagrangian as $L(\boldsymbol{B}, \theta, \lambda) = f(\boldsymbol{B}, \theta) + \lambda h_1(\boldsymbol{B})$, where $f(\boldsymbol{B}, \theta)$ is some loss such as the least square error in NOTEARS (X. Zheng, Aragam, et al. 2018) or the Kullback-Leibler divergence in DAG-GNN (Yu et al. 2019) with parameters θ, and λ is the Lagrange multiplier. Other causal structural learning algorithms (see e.g., Bühlmann et al. 2014; Chickering 2002; Kalisch and Bühlmann 2007; Ramsey et al. 2017; Shimizu et al. 2006; Spirtes, C. Glymour, et al. 2000; S. Zhu et al. 2020) can also be applied by formulating the corresponding score or loss function.

In order for $\boldsymbol{B}$ to satisfy structural constraints under decision-oriented CSL, such as $g_1(\boldsymbol{B})$ to $g_3(\boldsymbol{B})$ (see Section 5.3.1), it must satisfy: $h_2(\boldsymbol{B}) = \sum_{i=1}^3 g_i(\boldsymbol{B}) = 0$. As remarked earlier, more structural constraints can be added, and any added would be included in $h_2(\boldsymbol{B})$. Combining the two constraints above (h_1 and h_2) yields the following objective loss by an augmented Lagrangian (Cai et al. 2020),

$$L(\boldsymbol{B}, \theta) = f(\boldsymbol{B}, \theta) + \lambda_1 h_1(\boldsymbol{B}) + \lambda_2 h_2(\boldsymbol{B}) + c|h_1(\boldsymbol{B})|^2 + d|h_2(\boldsymbol{B})|^2,$$

where model parameter θ, λ_1 and λ_2 are Lagrange multipliers, and c and d are tuning parameters to ensure a hard constraint on h_1 and h_2.

5.3.3 Model Identifiablities. In the absence of further assumptions regarding the form of functions and/or noises, the model in (3) can only be identified up to MEC following the Markov and faithful assumptions (Peters, Mooij, et al. 2014; Spirtes, C. Glymour, et al. 2000). Below, we explore the conditions for the unique identifiability of the DAG and potential strategies for addressing scenarios involving the MEC. More specifically, a general causal DAG, $\mathcal{G}$, may not be identifiable from the distribution of $\boldsymbol{X}$. According to Pearl (2000), a DAG only encodes conditional independence relationships through the concept of d-separation. In general, several DAGs can encode the same conditional independence relationships, and such DAGs form a Markov equivalence class. Two DAGs belong to the same Markov equivalence class if and only if they have the same skeleton and the same v-structures (Kalisch and Bühlmann 2007). A Markov equivalence class of DAGs can be uniquely represented by a completed partially directed acyclic graph (CPDAG) (Spirtes, C. Glymour, et al. 2000), which is a graph that can contain both directed and undirected edges. A CPDAG satisfies the following: $X_i \leftrightarrow X_j$ in the CPDAG if the Markov equivalence class contains a DAG including $X_i \to X_j$, as well as another DAG including $X_j \to X_i$. The Markov equivalence class for a fixed CPDAG $\mathcal{C}$ is denoted by $\text{MEC}(\mathcal{C})$, which is a set containing all DAGs $\mathcal{G}$ that have the CPDAG structure $\mathcal{C}$. If we can obtain the true DAG from the data, we can simply treat it as a special case of the "MEC" containing only this DAG, i.e., $\text{MEC}(\mathcal{G}) = \{\mathcal{G}\}$.

Initially, we consolidate cases where the DAG is uniquely identifiable. In the context of the LSEM, when the noises $\boldsymbol{\epsilon}$ follow a Gaussian distribution, the resulting model corresponds to the standard linear-Gaussian model class, as investigated in Spirtes, C. Glymour, et al. (2000) and Peters, Janzing, et al. (2017). In instances where the noises $\boldsymbol{\epsilon}$ maintain equal variances, according to Peters and Bühlmann (2014), the DAG $\mathcal{G}$ can be uniquely identified

from observational data. Further, when the functions are linear but the noises are non-Gaussian, one can derive the LiNGAM as described in Shimizu et al. (2006), where the true DAG can be uniquely identified under certain favorable conditions. In addition, as cited in Rolland et al. (2022) and X. Zheng, Dan, et al. (2020), the nonlinear additive model can be identified from observational data. Another scenario of note arises when the corresponding MEC encompasses only one DAG, in which case the DAG can be inherently identified from observational data. Recent score-based causal discovery algorithms (Cai et al. 2020; Yu et al. 2019; X. Zheng, Aragam, et al. 2018; S. Zhu et al. 2020) typically take into account synthetic datasets generated from fully identifiable models, which provide practical relevance for evaluating the estimated graph in relation to the true DAG.

When the true DAG is not identifiable, a CPDAG uniquely symbolizes a MEC of DAGs that yield the same joint distribution of variables. This CPDAG can be inferred from observational data via a variety of causal discovery algorithms (see e.g., Bühlmann et al. 2014; Chickering 2002; N. Harris and Drton 2013; Kalisch and Bühlmann 2007; Ramsey et al. 2017; Shimizu et al. 2006; Spirtes, C. Glymour, et al. 2000; J. Zhang and Bareinboim 2018b). One feasible approach to dealing with MEC involves enumerating all DAGs in the MEC derived from a given CPDAG (Chakrabortty et al. 2018). It is conventional to encapsulate a range of potential effects or probabilities by their average or the minimum absolute value (Chakrabortty et al. 2018; Shi and L. Li 2021). However, such an approach typically proves computationally prohibitive for large graphs, necessitating computational shortcuts to acquire the causal effects or probabilities of causation without enumerating all DAGs in the MEC of the estimated CPDAG. With the additional identification constraints in the decision-oriented CSL, the size of the MEC is reduced, making it easier to achieve unique identification based on the observational data.

5.3.4 Decision-Oriented Causal Mediation Analysis. Causal mediation analysis holds significant importance in causal decision making, particularly due to its ability to interpret causal mechanisms through mediators. This analysis is challenging yet highly sought after, as it effectively bridges the gap between CSL and CEL. The integration of causal mediation analysis into decision-making processes enables a deeper understanding of how different variables and interventions interact and influence each other, leading to more informed and effective decisions. Another key motivation behind the use of causal mediation analysis is its role in CPL. By understanding the pathways through which causal effects are transmitted, policymakers and researchers can develop more nuanced and effective strategies. Identifying the causality among variables enables us to understand the key factors that influence the target variable, quantify the causal effect of an exposure on the outcome of interest, and use these effects to further guide downstream machine-learning tasks.

To visualize causes and counterfactuals, Pearl (2009) proposed to use the causal graphical model and the 'do-operator' to quantify the causal effects. A number of follow-up works (e.g., Chakrabortty et al. 2018; Maathuis et al. 2009; Nandy, Maathuis, et al. 2017) have been developed recently to estimate direct and indirect causal effects that are regulated by mediators in the linear SEM. These studies relied on the PC algorithm (Spirtes, C. Glymour, et al. 2000), which requires strong assumptions of graph sparsity and noise normality due to computational limits. To overcome these difficulties, Cai et al. (2020) proposed to leverage score-based CSL methods (e.g., Ramsey et al. 2017; Yu et al. 2019; X. Zheng, Aragam, et al. 2018; S. Zhu et al. 2020) with background causal knowledge to estimate mediation effects. In the following, we detail the analysis of causal effects (ANOCE) (Cai et al. 2020). Let A be the exposure/treatment, $\mathbf{M} = [M_1, M_2, \cdots, M_p]^\top$ be mediators with dimension p, and R be the outcome of interest. Suppose there exists a weighted DAG $\mathcal{G} = (\mathbf{Z}, B)$ that characterizes the causal relationship among $\mathbf{Z} = [A, \mathbf{M}^\top, R]^\top$, where the dimension of $\mathbf{Z}$ is $d = p + 2$. We next give the total effect (TE), the natural direct effect that is not mediated by mediators (DE), and the natural indirect effect that is regulated by mediators (IE) defined

in Pearl (2009).

$$TE = \partial E\{R|do(A=a)\}/\partial a = E\{R|do(A=a+1)\} - E\{R|do(A=a)\},$$
$$DE = E\{R|do(A=a+1, \mathbf{M}=\mathbf{m}^{(a)})\} - E\{R|do(A=a)\},$$
$$IE = E\{R|do(A=a, \mathbf{M}=\mathbf{m}^{(a+1)})\} - E\{R|do(A=a)\},$$

where $do(A=a)$ is a mathematical operator to simulate physical interventions that hold A constant as a while keeping the rest of the model unchanged, which corresponds to removing edges into A and replacing A by the constant a in $\mathcal{G}$. Here, $\mathbf{m}^{(a)}$ is the value of $\mathbf{M}$ if setting $do(A=a)$, and $\mathbf{m}^{(a+1)}$ is the value of $\mathbf{M}$ if setting $do(A=a+1)$. Refer to Pearl (2009) for more details of the 'do-operator'. First, we will define the natural direct effect of an individual mediator (DM).

$$DM_i = \Big[E\{M_i|do(A=a+1)\} - E\{M_i|do(A=a)\}\Big]$$
$$\times\Big[E\{R|do(A=a, M_i=m_i^{(a)}+1, \Omega_i=o_i^{(a)})\} - E\{R|do(A=a)\}\Big],$$

where $m_i^{(a)}$ is the value of M_i when setting $do(A=a)$, $\Omega_i = \mathbf{M} \setminus M_i$ is the set of mediators except M_i, and $o_i^{(a)}$ is the value of Ω_i when setting $do(A=a)$. The natural indirect effect for an individual mediator (IM) can be defined similarly.

$$IM_i = \Big[E\{M_i|do(A=a+1)\} - E\{M_i|do(A=a)\}\Big]$$
$$\times\Big[E\{R|do(A=a, M_i=m_i^{(a)}+1)\} - E\{R|do(A=a, M_i=m_i^{(a)}+1, \Omega_i=o_i^{(a)})\}\Big].$$

Based on the result $TE = DE + IE$ in Pearl (2009) and the above definitions, we summarize the defined causal effects and their relationship in Table 2 for the analysis of causal effects (ANOCE). Firstly, the causal effect of A on Y has two sources: the direct effect from A and the indirect effect via p mediators $\mathbf{M}$ $(M_1, \cdots, M_p)$. Next, the direct source has the degree of freedom ($d.f.$) as 1, while the indirect source has $d.f.$ as p from p mediators. Note the true $d.f.$ of the indirect effect may be smaller than p, since A may not be regulated by all mediators. Then, the causal effect for the direct source is the DE and for the indirect source is the IE, where the IE can be further decomposed into p DMs and each component corresponds to the natural direct effect for a specific mediator. The last row in the table shows that the DE and the IE compose the total effect TE with $d.f.$ as $p+1$. The ANOCE-CVAE learner (Cai et al. 2020) is a constrained CSL method by incorporating a novel identification constraint that specifies the temporal causal relationship of variables. The above decision-oriented causal mediation analysis involves causal structures with possibly multiple mediators under the structural equation model. In Section 6.2.3, we will detail the semi-parametric efficient estimation of mediation effects with a single mediator in the framework of causal effect learning.

5.4 Decision-Oriented CSL for Paradigm 2+

Recent advances in causal discovery for time series data have significantly pushed the boundaries of this field to non-i.i.d. settings. Traditional approaches such as Granger causality (Granger 1969) laid the foundation for statistical time series analysis but fall short when handling nonlinearity, high dimensionality, and complex temporal dependencies. More recent developments, such as the PCMCI algorithm (Runge et al. 2019), build upon rigorous theoretical foundations (Runge 2018) to effectively disentangle causal relationships in the presence of autocorrelation and time-lagged confounding. These methods leverage conditional independence testing combined with constraint-based structure learning, providing improved scalability and reliability in multivariate settings. Comprehensive evaluations of time-series causal discovery methods, such as the survey by Assaad et al. (2022), highlight the trade-offs among existing techniques in terms of their statistical power, robustness

Table 2. Table of Analysis of Causal Effects (ANOCE Table).

Source	Degree of freedom	Causal effects
Direct effect from A	1	DE
Indirect effect via M	p	IE
$\begin{cases} M_1 \\ M_2 \\ \vdots \\ M_p \end{cases}$	$\begin{cases} 1 \\ 1 \\ \vdots \\ 1 \end{cases}$	$\begin{cases} DM_1 \\ DM_2 \\ \vdots \\ DM_p \end{cases}$
Total	$1+p$	TE

to noise, and computational scalability. Their work serves as a useful taxonomy and benchmarking framework for understanding the empirical behavior of algorithms across various simulated and real-world time series datasets. In parallel, machine learning approaches have emerged as powerful tools for nonlinear causal inference in time series. For instance, Gaussian process-based methods (Lopez-Paz and Schölkopf 2017) offer a flexible nonparametric framework for modeling functional relationships, while recent deep learning-based frameworks such as Temporal Causal Discovery Framework (TCDF) exploit attention mechanisms to capture temporal dependencies and infer cause-effect links in high-dimensional sequences (Nauta et al. 2019). The integration of information-theoretic concepts like transfer entropy into deep models (Tank et al. 2021) further enriches the toolbox for analyzing dynamic systems by quantifying directed dependencies without strong parametric assumptions.

These innovations have not only improved the accuracy and interpretability of causal analysis in temporal settings but also broadened the scope of applications to domains such as neuroscience, climate science, and finance. However, to the best of our knowledge, these methods predominantly operate within observational settings and do not explicitly incorporate treatments or decisions as central components of the causal model. Consequently, decision-oriented causal discovery for Paradigms 2+ remains a largely open and unexplored frontier in the current literature.

6 Causal Effect Learning

This section aims to provide a detailed introduction to causal effect learning. We categorize CEL into three groups based on the underlying causal structure: CEL with 1) i.i.d. data (ATHEY et al. 2019; Künzel et al. 2019), 2) Markov transition state (N. Jiang and L. Li 2016; Kallus and Uehara 2022; Q. Liu et al. 2018), and 3) panel data (Lechner et al. 2011; Viviano and Bradic 2022). Our aim is to provide a systematic review of estimation techniques for the average treatment effect, the heterogeneous treatment effect, and the mediation effect under the above three scenarios.

In the context of CEL, our aim is to answer the following question:

What is the causal effect of some intervention/treatment/policy?
If it's well-defined, then how can we quantify it stably and efficiently?

This question mainly concludes the three main tasks in the realm of CEL: identification, estimation, and inference.

Let us take the MIMIC-III data as an example. Once the causal structure of the data (including all potential confounders and mediators) has been determined, the next step is to quantify the effect of intravenous input on the mortality status of the patients. This problem involves identification under reasonable assumptions (such as no latent confounders), estimating the average and personalized effects of IV input on mortality status, and

ultimately finding the optimal policy to tailor personalized medical treatments to minimize the overall mortality rate.

Among these three tasks, we are interested in different causal estimands: Average Treatment Effect (ATE), Heterogeneous Treatment Effect (HTE), and mediation effect. In the next subsections, we will mix the three main tasks and three main estimands together to conduct a comprehensive review according to the data structure of different paradigms.

Table 3. Summary of Causal Effect Learning Literature

Application	Method Type	Model & Paper
Single-stage ATE	DM	G. W. Imbens and Wooldridge (2009)
	IPW	Hirano et al. (2003) and Rosenbaum and Rubin (1983)
	DR	Bang and Robins (2005) and Chernozhukov, Chetverikov, et al. (2018)
Single-stage HTE	Meta-learners	S-, T-, and X-learner (Künzel et al. 2019)
	R-learner	Nie and Wager (2021)
	DR-learner	Kennedy et al. (2020)
	Causal Forest	Wager and Athey (2018)
	Dragonnet	Shi, Blei, et al. (2019)
	review paper	Curth and Schaar (2021)
Mediation Analysis	Parametric approaches	causal steps (Baron and Kenny 1986)
		difference in coefficients (MacKinnon and Dwyer 1993)
		product of coefficients (Alwin and R. M. Hauser 1975)
	DM	Imai, Keele, and Tingley (2010)
	IPW	Hong et al. (2010)
	DR	Tchetgen and Shpitser (2012)
Panel Data	DiD	Lechner et al. (2011)
	Synthetic control	Abadie and Gardeazabal (2003) and K. T. Li (2020)
	Matrix completion	Athey, Bayati, et al. (2021)
	Synthetic DiD	Arkhangelsky, Athey, et al. (2021)

6.1 Why Causal Effect Learning is Needed for Causal Decision Making

CEL aims to accurately quantify the causal effect of some intervention/policy on a group of units. As an intermediate stage of causal decision making, CEL plays an important role in conducting a primary analysis of a given causal diagram, as well as providing the necessary information for post-stage policy learning and decision making. The internal connections are multi-fold and detailed below.

First, CEL provides *primary insights* for decision making. To answer the question of "which policy yields the highest reward or desired outcome within a given population", a fundamental prerequisite is comprehending how a given policy impacts distinct units within that population in a heterogeneous manner. This challenge, often encapsulated in the estimation of HTE, aligns perfectly with the domain of CEL. In experimental design, A/B testing is a widely used method in industry to measure the effectiveness of changes or interventions. In observational studies, HTE estimation (such as $\tau(s) = \mathbb{E}[R(1) - R(0) \mid S = s]$ in a binary action space) can be directly applied to decision making by selecting the action $\mathbf{1}\{\tau(s) > 0\}$. In general, CEL with observational data provides valuable insights into the effectiveness of specific treatments in a more cost-efficient manner.

Second, CEL offers *a systematic identification framework* that supports the *validity* of decision making. Firstly, while not always explicitly stated, most decision-making methods in RL rely on certain causal assumptions, such as SUTVA and NUC, as outlined in Section 3.4. These assumptions, though sometimes restrictive, ensure the identifiability of specific value functions, which is essential to conduct valid policy learning. Secondly, in more complex scenarios with interference issues (Sävje et al. 2021) or unmeasured confounders (L. Wang and Tchetgen Tchetgen 2018), CEL incorporates techniques such as instrumental variables or specifying interference structures (see Section 9), supporting reliable decision making based on effect estimates. For example, when assessing whether smoking increases the risk of lung cancer, genetic predisposition serves as a confounder, as it can causally influence both the likelihood of smoking and the risk of developing lung cancer. Properly accounting for this unseen factor is essential to avoid misleading conclusions. In summary, CEL acts as a "safeguard", formalizing the identification framework to ensure that the effect learning remains estimable and that the subsequent decision making is valid.

Third, CEL *filters out ineffective treatment options with confidence* for better decision making. In addition to providing value estimates, CEL also serves as a platform for statistical inference on the causal effects of interest (Athey, G. W. Imbens, and Wager 2018; Benkeser et al. 2017; Mealli et al. 2011). For example, consider a decision-making problem in recommending personalized treatments for patients in a clinical trial. Although point estimates of the value function for different treatment options reflect expected outcomes, accurate statistical inference goes further by quantifying uncertainty, thereby helping determine the need for quasi-control treatment options and simplifying the decision-making process with greater confidence.

6.2 CEL in Single Stage (Paradigm 1)

Over the past decades, there has been extensive study on conducting causal inference in the classical single-stage setup, where only one state-action-reward tuple is observed for an individual. Next, we will detail some representative approaches according to the specific tasks they are dealing with: i) ATE, ii) HTE, and iii) mediation effect.

6.2.1 ATE. Based on the "big three assumptions" of causal inference (see Assumption 3.1-3.3 in Section 3.4), there are representative ATE estimation methods in literature, commonly referred to as the direct (outcome regression) estimator (DM), inverse probability weighting (IPW) estimator, and doubly robust (DR) estimator. As we move to later sections, we will see that the concept of DR estimation is widely applied in various effect estimations, including HTE and mediation analysis, across both single-stage and infinite-horizon settings.

The intricacy of causal inference manifests prominently in the challenge of counterfactual estimation. It is inherently impossible to directly observe the outcomes users would obtain had they chosen differently at the time of treatment assignment. However, failing to observe counterfactuals doesn't mean that we cannot estimate/infer interesting quantities under some reasonable assumptions, which are detailed in Section 3.4. When SUTVA, NUC, and Positivity assumptions hold, the potential outcomes can be rewritten as $\mathbb{E}[R(a)|\boldsymbol{S}=\boldsymbol{s}] = \mathbb{E}[R|\boldsymbol{S}=\boldsymbol{s}, A=a]$, where the right-hand side is entirely estimable from observational data. Therefore, the most intuitive way is to estimate the counterfactual part via a regression model. This yields the first estimator, known as the direct method, as outlined below:

$$\widehat{\mathrm{ATE}}_{\mathrm{DM}} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}(\boldsymbol{S_i}, 1) - \hat{\mu}(\boldsymbol{S_i}, 0)\}, \tag{7}$$

where $\hat{\mu}(\boldsymbol{s}, a)$ is the estimated outcome regression model for $\mathbb{E}[R|\boldsymbol{S}=\boldsymbol{s}, A=a]$.

The second type of estimator is called the inverse probability weighting (IPW) estimator, or importance sampling (IS) estimator in RL literature. Define propensity score $\mathbb{P}(A=1|\boldsymbol{S})$ as the probability of receivting treatmente $A=1$. IPW estimator uses propensity scores to reweight observations, balancing the distribution of

covariates between the treated and control groups by mimicing a randomized experiment.

$$\widehat{\mathrm{ATE}}_{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i R_i}{\hat{\pi}(\boldsymbol{S_i})} - \frac{(1-A_i)R_i}{1-\hat{\pi}(\boldsymbol{S_i})}\right\}. \tag{8}$$

By combining both estimators, the DR estimator (or augmented IPW, i.e. AIPW) is consistent as long as either the outcome regression model or the propensity score model is correctly specified.

$$\widehat{\mathrm{ATE}}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}(\boldsymbol{S_i},1) - \hat{\mu}(\boldsymbol{S_i},0) + \frac{A_i(R_i - \hat{\mu}(\boldsymbol{S_i},1))}{\hat{\pi}(\boldsymbol{S_i})} - \frac{(1-A_i)(R_i - \hat{\mu}(\boldsymbol{S_i},0))}{1-\hat{\pi}(\boldsymbol{S_i})}\right\}. \tag{9}$$

Under certain mild entropy conditions or through sample splitting, the DR estimator is also a semi-parametrically efficient estimator when the convergence rate of both $\hat{\mu}$ and $\hat{\pi}$ are at least $o(n^{-1/4})$. The key insight behind this robustness, i.e., the orthogonalization of nuisance parameters and the use of sample splitting to avoid bias from overfitting, was later formalized and extended in the Double Machine Learning (DML) framework by Chernozhukov, Chetverikov, et al. (2018).

Although the AIPW estimator guarantees double robustness, it may still result in a poor estimator when both the propensity score and outcome regression models are not correctly specified. In addressing this challenge, an alternative line of research has emerged, focusing on reducing estimation bias via the optimization of model parameters (Vermeulen and Vansteelandt 2016; S. Yang, J. K. Kim, et al. 2020). Additionally, to address estimates that fall outside the admissible parameter range (e.g., a mortality rate outside $[0, 1]$), Targeted Maximum Likelihood Estimation (TMLE) (Gruber and M. J. v. d. Laan 2010; Gruber and Van Der Laan 2012) was developed to incrementally adjust the estimator while maintaining the double robustness of AIPW method.

In the ideal scenario where the treatment and control groups share similar covariate distributions, the aforementioned estimators are expected to perform well. However, when there is strong selection bias, matching techniques (Abadie and G. W. Imbens 2011, 2006, 2008; Caliendo and Kopeinig 2008; Heckman et al. 1998) offer a valuable avenue to improve the performance of the estimation. As the gold standard of causal inference, randomized experiments impose fewer assumptions for identification and estimation. Therefore, the fundamental idea of matching is a way to find the closest "randomized experiment" hidden inside the observational study, to adjust for covariate imbalances between groups. Under a certain distance metric, one of the most intuitive ways is to select the top k nearest neighbors for each unit, and conduct an average to estimate the corresponding counterfactuals (Abadie and G. W. Imbens 2008).

However, in scenarios with a relatively large number of covariates, traditional distance metrics for neighbor selection may encounter challenges due to the curse of dimensionality. To address this, the propensity score (Abadie and G. W. Imbens 2016; Austin 2008; Rosenbaum and Rubin 1983) and the prognostic score (Hansen 2008) are commonly used as two representative balancing scores to reduce dimensionality when adjusting for covariate imbalance. Later, double score (DS) matching (Antonelli et al. 2018; Leacy and Stuart 2014; S. Yang and Y. Zhang 2023) was proposed to jointly combine the above two scores, which further improves the matching performance. This method is double robust in the sense that the DS matching estimator remains consistent for ATE if either the propensity score or the prognostic score is correctly specified. For further practical insights into matching methodologies, refer to Y. Zhang, S. Yang, et al. (2022).

Overall, the methods of ATE estimation have been thoroughly studied in literature over the past decades. For other related review papers, we refer to Yao et al. (2021) under the potential outcome's framework and Pearl (2009) under SCM. In most decision-making contexts involving a specific population, ATE estimation plays a crucial role in quantifying the overall impact of different decision rules. This is applied across various domains, including, but not limited to, assessing the effectiveness of advertising campaigns (Farahat and Bailey 2012), labor market interventions in public policy (Dehejia and Wahba 1999), and epidemiology (M. A. Hernán and Robins 2006). In

summary, ATE estimation serves as a fundamental tool for determining the population or sub-population-level effects of different treatments.

6.2.2 HTE. In real applications, our focus usually extends beyond the average treatment effect at the population level; rather, the estimation of personalized treatment effects for individuals or within specific subgroups often draws our interest. For example, with MIMIC-III data, our ultimate goal is to figure out the optimal IV input strategy for each patient, which often starts with understanding the personalized treatment effect at an individual level.

Existing work in single-stage HTE estimation starts from meta-learners (Künzel et al. 2019) and subsequently extends to more comprehensive approaches, which either demonstrate improved theoretical properties in statistical inference (Kennedy 2023) or exhibit enhanced performance in specific settings (Nie and Wager 2021; Shi, Blei, et al. 2019). We will detail some representative methods below.

The first type of learners are meta-learners, which consist of S-learner, T-learner, and X-learner. Under the SUTVA, NUC, and positivity assumptions, we have

$$\tau(\boldsymbol{s}) = \mathbb{E}[R(1) - R(0)|\boldsymbol{S} = \boldsymbol{s}] = \mathbb{E}[R|\boldsymbol{S} = \boldsymbol{s}, A = 1] - \mathbb{E}[R|\boldsymbol{S} = \boldsymbol{s}, A = 0].$$

If we estimate $\mathbb{E}[R|\boldsymbol{S} = \boldsymbol{s}, A = 1]$ and $\mathbb{E}[R|\boldsymbol{S} = \boldsymbol{s}, A = 0]$ together by fitting $R \sim (\boldsymbol{S}, A)$, we obtain what is known as the S-learner. Conversely, if we divide the data into treated and control groups, and fit $\mathbb{E}[R|\boldsymbol{S} = \boldsymbol{s}, A = 1]$ and $\mathbb{E}[R|\boldsymbol{S} = \boldsymbol{s}, A = 0]$ separately with two independent models, this gives rise to the T-learner. While both learners are straightforward to implement, the S-learner tends to exhibit slightly better performance when the treated group and control group share a similar reward structure. Conversely, T-learner may be preferable due to its ability to differentiate action A from all other covariates X in reward modeling. This distinction prevents the risk of neglecting the "action" among other covariates, which could occur with the S-learner. Based on the two base learners, the X-learner was proposed by Künzel et al. (2019), which shows more favorable performance, especially when dealing with sample size imbalance between the treatment and the control group, or when the separate parts of the X-learner can exploit the structural properties (such as smoothness or sparsity) of the reward and treatment effect functions.

Later, several additional learners were proposed, including the R-learner (Nie and Wager 2021), the DR-learner, and the Lp-R-learner (Kennedy 2023), all following a two-step approach and demonstrating promising theoretical results. The concept of R-learner originated from P. M. Robinson (1988) in 1988 and was formalized by Nie and Wager (2021) in 2021. R-learner, which stands for "residual" learner, is a two-step methodology that involves regressing reward residuals on propensity score residuals, which is able to adapt to various modeling needs and ensure a quasi-oracle property with penalized kernel regression. DR-learner, introduced by Kennedy (2023), integrates insights from the DR estimator to construct an HTE estimator at the first stage, followed by regression on pseudo outcomes to obtain the final learner. In the same paper, the Lp-R-learner combines residual regression with local polynomial adaptation, employing cross-fitting to relax conditions for achieving the oracle convergence rate. Despite providing promising theoretical results, this algorithm may incur computational intensity when applying local polynomial regressions to a large degree.

Recently, a new stream of work incorporates neural networks in HTE estimation to provide potentially more flexible modeling choices. The majority of existing work shares a similar two-step pattern in HTE estimation: In Step 1, the nuisance functions (including propensity score and outcome models) are fitted via some NN-based methods; in Step 2, fitted nuisance functions are combined to estimate the HTE via some downstream estimating equation (e.g., plug-in estimator, IPW estimator, or DR estimator). Notably, Shi, Blei, et al. (2019) proposed a novel neural network architecture based on the sufficiency of the propensity score for causal estimation in Step 1, and a regularization procedure in Step 2 to optimize nonparametric performance. Similar neural network-based approaches are explored in related works such as Hassanpour and Greiner (2019), Johansson et al. (2016), and

Shalit et al. (2017). For a comprehensive review and comparison of these different approaches, we refer to Curth and Schaar (2021) for a more detailed review.

In many application scenarios, effect learning is often a crucial pre-step before making decisions. The relationship between HTE and decision-making can be as simple as $\mathbf{1}\{\tau(\boldsymbol{s}) > 0\}$, or can be adapted to more realistic concerns such as resource or budget constraints. For example, decisions can be made by selecting treatments for patients whose predicted treatment effect exceeds a decision threshold (Dorresteijn et al. 2011) in clinical trials. Alternatively, decision-making can involve a more complex function of HTE, incorporating factors such as costs and pricing associated with actions (Miller and Hosanagar 2020), or serve as an intermediate step that feeds into downstream optimization tasks under resource constraints (Qiu, Carone, and Luedtke 2022). The flexibility of the HTE methods introduced above allows decision-makers to select the most appropriate approach based on their specific needs.

6.2.3 Mediation Effect. Mediators are variables that are causally affected by action A and, in turn, influence the reward modeling of R. They create an additional causal pathway from action to reward, which is often considered when analyzing complex causal relationships. In Section 5.3.4, we discuss several key methods in CSL to identify mediators in causal graphs, with particular attention to recent advances in score-based approaches (Cai et al. 2020) that address mediator identification and effect learning simultaneously. While it is great to kill two birds with one stone, this type of approach may suffer from potential limitations in modeling, such as linearity.

In this section, we shift our focus to CEL, specifically in the context of a single mediator when the causal structure is already known. Mediation effect estimation approaches can be roughly divided into three categories: (1) classical approaches based on parametric modeling, (2) non-parametric and semi-parametric causal mediation analysis, and (3) later extensions. Compared to the methods discussed in Section 5.3.4, the more recent techniques emphasize flexible modeling, which can be particularly advantageous when focusing solely on effect learning with a known mediation structure.

Causal mediation analysis is well-developed in single-stage settings. Recently, a growing body of work has focused on extending these approaches to MDPs and other data structures, such as DTR. Here, we will briefly summarize the main approaches in paradigm 1 and refer to some review papers for further reading.

We will start by introducing the classical approaches (MacKinnon, Fairchild, et al. 2007). In the presence of a mediator M, the causal relationship between action A and reward R is illustrated in Figure 6. Classical approaches focus on decomposing the strength of causal paths using three parametric models.

$$\begin{aligned} R &= \beta_1 + cA + \epsilon_1 \\ R &= \beta_2 + c'A + bM + \epsilon_2 \\ M &= \beta_3 + aA + \epsilon_3, \end{aligned} \tag{10}$$

where the coefficients (a, b, c, c') correspond to the strength of the causal relationships depicted in Figure 6. There are three main approaches based on these equations, which are (i) causal steps (Baron and Kenny 1986), (ii) difference in coefficients (MacKinnon and Dwyer 1993), and (iii) product of coefficients (Alwin and R. M. Hauser 1975). The causal step approach is a four-step regression-based procedure to decompose the significance and strength of different paths in Figure 6. The difference in coefficients approach approximates the mediated effect by calculating $\hat{c} - \hat{c}'$, while the product of the coefficients approach estimates the mediated effect by $\hat{a}\hat{b}$. The last two estimators are equivalent when modeling with linear regression. All three approaches are widely used in practical applications due to their simplicity and interpretability. Later on, there are some follow-up reviews under the linear structural equation model (LSEM) (Bollen 1987; Hayes 2017; Imai, Keele, and Yamamoto 2010; MacKinnon, Lockwood, Hoffman, et al. 2002; Pearl 2022) that allows for measuring the causal relationships between multiple variables in a more flexible way. However, these approaches may also suffer from the drawbacks of parametric assumptions, such as linearity.

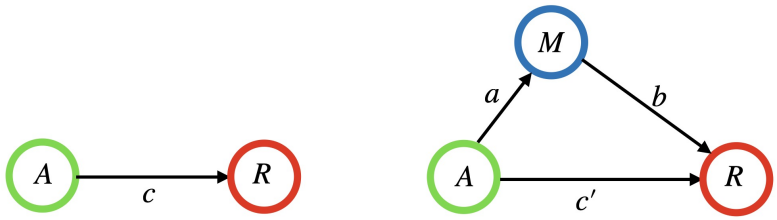


Fig. 6. Mediation effect. The left panel illustrates the total causal effect from A to R, while the right panel shows its decomposition into the direct effect (c') and the indirect effect via the mediator, represented by paths a and b.

The second type of estimator is based on more recent studies on causal mediation analysis under the potential outcomes framework (Imai, Keele, and Tingley 2010). In recent years, extensive work has focused on non-parametric identification and non(semi)-parametric estimation and inference of mediated direct and indirect effects (Tchetgen and Shpitser 2012). Similar to ATE estimation, researchers have proposed corresponding versions of DM (Imai, Keele, and Tingley 2010), IPW (Hong et al. 2010), and DR (Tchetgen and Shpitser 2012) estimators for estimating direct and indirect effects in mediation analysis. Notably, the DR mediation effect estimands proposed by Tchetgen and Shpitser (2012) achieve semi-parametric efficiency.

The third stream of work focuses on relaxing certain NUC assumptions or modeling requirements. For instance, some studies handle binary mediators using principal stratification (Gallop et al. 2009; Rubin 2004; T. J. VanderWeele 2008). Others relax the linear assumptions in LSEM by employing alternative regression models (MacKinnon, Lockwood, Brown, et al. 2007) or by incorporating exposure-by-covariate and mediator-by-covariate interactions (Hayes 2017). Additionally, some work allows for the presence of specific types of confounders (T. VanderWeele 2015; T. J. VanderWeele and Vansteelandt 2009). To accommodate the flexibility required for multi-stage decision making, recent studies have also addressed mediation effect estimation in reinforcement learning (paradigm 2) (Ge et al. 2023) and dynamic treatment regimes (paradigm 3) (D. L. Roth and MacKinnon 2013; Selig and Preacher 2009; W. Zheng and M. v. d. Laan 2017). We will further elaborate on their connections within our proposed framework in Sections 6.3–6.4.

6.3 CEL Under MDP (Paradigm 2)

In some real cases, researchers may encounter longitudinal data with long horizons or even infinite horizons. For example, in clinical trials, the doctor will periodically check the health status of patients to provide them with personalized treatment each time they visit. Under this scenario, we aim to estimate the long-term causal effect of taking a specific treatment across all stages. Under the Markovian assumption, this problem is referred to as causal effect estimation within an MDP framework, as detailed in Definition 3.7.

Unlike the single-stage setting where much work focuses on estimating the difference in potential outcomes under $A = 1$ and $A = 0$, the definition of causal effect becomes more general in multi-stage settings after introducing the concept of *policy*. As defined in Definition 3.11, a policy $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping from state to action space that quantifies the treatment assignment strategy for different actions in $\mathcal{A}$. The causal effect estimation problem is thus generalized to estimating the state-value function $V^{\pi}(s) = \sum_{t'=t}^{T} \mathbb{E}^{\pi}[\gamma^{t-t'} R_{t'} | S_t = s]$, a discounted cumulative reward aggregated under policy π. Following similar logic in single stage, we can still define HTE and ATE under MDP or any other multi-stage settings as the difference in the value function under two policies (π and π_0), i.e.,

$$\text{HTE}(s) = V^{\pi}(s) - V^{\pi_0}(s), \quad \text{and} \quad \text{ATE} = \mathbb{E}_{s\sim\mathcal{S}}\Big[V^{\pi}(s) - V^{\pi_0}(s)\Big].$$

While we can naively estimate the quantities using techniques in single-stage CEL by regarding π and π_0 as two treatments, this approach overlooks the unique Markovian structure in state transitions, resulting in less efficient estimates. Instead, by leveraging the sequential decision-making structure and the Markovian assumptions, we can derive estimators that converge much faster.

By involving the definition of *policy*, the problem of HTE and ATE estimation can be regarded as a direct byproduct of conducting policy evaluation on the value function $V^{\pi}(s)$ (Shi, X. Wang, et al. 2023). In the RL literature, this is widely known as Off-Policy Evaluation (OPE). As this part highly overlaps with CPL in paradigm 2, we will leave the main discussion to Section 7.1.

6.4 CEL in Panel Data (Paradigm 3)

In addition to the MDP structure in paradigm 2, panel data represents another common multi-stage setting for causal effect estimation, which naturally aligns with paradigm 3 of our framework. Given its prominence as a subfield in the causal inference literature, we next review several widely used methodologies for panel data analysis.

Panel data analysis examines longitudinal data collected over time from the same individuals, companies, or entities (known as *panels*), often under varying treatment conditions. This approach is commonly used by governments and organizations to assess the long-term effects of policies on outcomes such as income, health, and education. By leveraging longitudinal data, panel analysis supports informed decision making and provides valuable insights into the lasting impacts of policy interventions. The traditional literature in panel data analysis primarily focuses on estimating the Average Treatment Effect on the Treated (ATT), defined as the expected difference in outcomes between treated and control units:

$$\mathbb{E}[R_{i,t}(1) - R_{i,t}(0)|G_i = 1], \tag{11}$$

where $G_i \in \{0, 1\}$ indicates whether unit i is in the treatment group ($G_i = 1$) or the control group ($G_i = 0$). Different from single-stage ATE/HTE estimation, panel data analysis aims to quantify the change of causal effect over time. Based on SUTVA, a classical assumption in causal inference, ATT can be identified from observed data. Since $R_{i,t}$ can be observed to unbiasedly estimate $\mathbb{E}[R_t(1)|G_i = 1]$, the main challenge of panel data analysis is to impute the missing values for $R_{i,t}(0)$ for treated units. That is, we would like to answer the question of *"What would happen to the treated units if they were exposed to control back to the treatment time?"*

Since the true answer is unobservable to us, we need to rely on additional assumptions to leverage existing information, particularly control units, for counterfactual estimation. To address this problem, there are two main streams of work in the literature: Difference-in-Difference (DiD) (Lechner et al. 2011) and Synthetic Control (SC) (Abadie and Gardeazabal 2003; K. T. Li 2020). DiD approach relies on the parallel trend assumption, where we learn the change of causal effect over time for control units and apply them to treated units for counterfactual estimation. Conversely, SC approximates each treated unit with a weighted combination of controls so as to borrow this weighted information for estimation. To be clearer, DiD borrows the information of control units over time and inherits it to treated units, while SC borrows the information of pre-treated stage over units and inherits it to post-treatment time. Due to the difference in estimation strategy, DiD is often used when we are willing to assume the parallel trend assumption, while SC is often applied to cases where only a few units are exposed to treatment.

Later, Athey, Bayati, et al. (2021) proposed a unified approach to integrate DiD and SC using matrix completion. Unlike DiD and SC, which rely on specific parallel or orthogonal assumptions, Athey, Bayati, et al. (2021) reframed causal effect estimation as a missing data imputation problem, assuming a low-rank structure to estimate counterfactuals in $\boldsymbol{R}(0)$. Recent advancements include but not limited to (1) R-DiD (Nie, C. Lu, et al. 2021), which extends classical DiD by relaxing linear functional assumptions to accommodate more flexible estimands, (2) Synthetic DiD (SDiD) (Arkhangelsky, Athey, et al. 2021), combining the benefits of DiD and SC by

re-weighting and matching pre-exposure trends to mitigate parallel assumptions while remaining invariant to additive unit-level shifts, and (3) Synthetic Learner (Viviano and Bradic 2023), an ensemble method enhancing precision through model-free inference.

This field is rapidly evolving, offering greater flexibility in modeling choices and relaxing assumptions. Classical literature in panel data analysis primarily emphasizes effect estimation under relatively simple decision choices, often by examining patterns of change before and after treatment assignment, with less focus on directly modeling policy learning. Recent work, such as K. Harris et al. (2024), introduces a strategy-proof framework for policy learning that maps pre-treatment outcomes to various intervention choices. This advancement has helped define policy learning explicitly within the context of panel data. For an in-depth overview, see the recent review by Hsiao (2022) and Arkhangelsky and G. Imbens (2024).

7 Offline Causal Policy Learning

This section presents policy evaluation and optimization methods for offline/off-policy settings (i.e., Paradigms 1-3). In contrast to online policy learning, additional data collection is infeasible in the offline setting, resulting in distribution shifts across multiple dimensions—particularly in actions and states—which poses a critical challenge. These shifts 1) introduce selection bias that necessitates causal adjustments as employed in CEL, and 2) increase uncertainty in policy evaluation and hence optimization, requiring a pessimistic or penalty-based approach to avoid over-optimization. In particular, because the data are observational and action-dependent, the outcomes of unchosen actions are systematically missing, and naïvely optimizing against observed rewards can lead to severe bias and unsafe policies. As a result, offline RL is fundamentally a causal problem: learning or evaluating a target policy requires reasoning about counterfactual outcomes under interventions that were not realized in the logged data. Core challenges in offline RL—such as confounding, distributional shift between behavior and target policies, and extrapolation beyond the support of observed actions—are naturally framed in terms of causal identifiability and counterfactual inference. While many classical offline RL methods are not explicitly formulated within a causal inference framework, they implicitly rely on causal assumptions analogous to ignorability, positivity, and consistency. We connect our review of related methods to the corresponding parts in CEL and CSL.

We begin with formal definitions of the tasks in CPL. To illustrate, we consider Paradigm 1, where the observed data consists of n data points $\{(S_i, A_i, R_i)\}_{1\le i\le n}$, collected by following a stationary *behavior policy* π_b. The two tasks in offline CPL are defined as:

- **Off-Policy Evaluation (OPE)**: The goal of OPE is to estimate the goodness of a given *target policy* π, which is typically evaluated by the integrated value $\eta^\pi = \mathbb{E}_{s\sim\mathbb{G}} V^\pi(s)$ with respect to some state distribution $\mathbb{G}$.
- **Off-Policy Optimization (OPO)**: The goal of OPO is to solve the optimal policy π^*, or in other words, to learn a policy $\hat{\pi}$ so as to minimize the regret $\eta^{\pi^*} - \eta^{\hat{\pi}}$.

7.1 Offline Policy Evaluation

OPE focuses on estimating the expected reward of an evaluation policy using historical data generated by a different behavior policy. This is particularly valuable in offline RL settings, where experimenting with policies is not possible due to ethical, financial, or safety concerns. OPE methods have gained importance across fields, including healthcare, education, and recommendation systems, where reliable evaluation of new policies without online testing is critical.

Model-based estimators. Model-based OPE (Paduraru 2013; Yin and Y.-X. Wang 2020) approaches estimate state transition and reward functions directly from data, which can then be used to simulate trajectories and estimate policy value. These methods achieve asymptotic efficiency in discrete MDPs. Such estimators often leverage probabilistic neural networks to model transitions, improving performance in complex continuous

Table 4. Summary of Offline CPL Literature

Application	Method Type	Paper
OPE	Model-based	Le et al. (2019), Paduraru (2013), Uehara, J. Huang, et al. (2020), and Yin and Y.-X. Wang (2020)
	Model-free	Dai et al. (2020), N. Jiang and L. Li (2016), Kallus and Uehara (2022), Q. Liu et al. (2018), Shi, Wan, Chernozhukov, et al. (2021), and J. Zhu et al. (2024)
OPO	Model-free Value-Based	Ernst et al. (2005), Liang et al. (2018), Murphy (2003), Robins, D. Lin, et al. (2004), Schulte et al. (2014), Shi, Fan, et al. (2018), Shi, S. Luo, et al. (2024), R. Song, W. Wang, et al. (2015), and W. Zhu et al. (2019)
	Model-based	Chua et al. (2018), Deisenroth and Rasmussen (2011), Levine, Finn, et al. (2016), Nagabandi et al. (2018), and Sutton (1991)
	Model-free Policy-based	Kitagawa and Tetenov (2018), Y. Liu et al. (2018), R. Song, M. Kosorok, et al. (2015), Tschernutter et al. (2022), Y. Zhang, Laber, et al. (2018), and Zhao, D. Zeng, et al. (2012)
	Pessimism / Penalty	Fujimoto et al. (2019), Jaques et al. (2019), Kumar et al. (2019), P. Liao et al. (2022), Rashidinejad et al. (2021), Siegel et al. (2020), and Y. Zhou et al. (2023)

control tasks. Although model-based methods allow for easier parameter tuning, particularly through supervised learning techniques, they can struggle in high-dimensional settings where modeling state transitions becomes more complex than direct estimation of value functions.

Model-free estimators. To adjust for selection bias caused by the distribution shift in offline datasets, the most popular methods include DM, IPW, and DR estimators. The classic forms of these methods can be derived from the following relationship:

$$\eta(\pi) = \mathbb{E}_{a\sim\pi(\cdot|S),s\sim p(S)}R(a) = \mathbb{E}_{a\sim\pi(S),s\sim p(S)}\Big[\mathbb{E}\big\{R(a)|S=s\big\}\Big] \tag{12}$$

$$= \mathbb{E}_{a\sim b(S),s\sim p(S)}\frac{\pi(a|s)}{b(a|s)}R(a) \tag{13}$$

$$= \mathbb{E}_{a\sim b(S),s\sim p(S)}\frac{\pi(a|s)}{b(a|s)}\Big[R(a)-V(s)\Big] + \mathbb{E}_{s\sim p(S)}V(s). \tag{14}$$

Specifically, by replacing $\mathbb{E}\big\{R(a)|S=s\big\} = Q(s,a)$ in (12) with its estimator, we obtain the direct method estimator; by replacing the expectation over $R(a)$ in (13) with a sample average of the observed rewards under action a, we obtain the IPW estimator; and finally we can combine these two approaches in (14) to derive the DR estimator. Notably, it is easy to see that these three methods are direct extensions of their counterparts in CEL (Equations (7), (8) and (9)), by taking additional expectation over the state and action distributions. Similar to the argument in CEL, these methods effectively employ different ways to adjust for the confounding effect from s, by either removing its imbalance across actions or its impact on the rewards.

Extensions to paradigms 2 and 3. These methods can all be extended to more complicated settings in Paradigms 2 and 3, by additionally accounting for the dependency over decision points. To simplify the problem in Paradigm 2, we can utilize the *recursive* or *iterative* structure. Take the direct method as an example, where as long as we can obtain an estimate of the Q-function, we can directly take its expectation to calculate the value as in (12). To estimate the Q-function, we introduce two prominent approaches. The most straightforward method, Fitted-Q Evaluation (FQE) (Le et al. 2019), leverages the Bellman Optimality Equation (1) which characterizes

the sequential dependency structure. FQE solves the loss function corresponding to (1) until we converge to a final value function estimator. Another method is Minimax Q-Learning (Uehara, J. Huang, et al. 2020), which enhances Q-function evaluation by framing it as a competition between two components: the Q-function itself and a discriminator function. This method leverages the Bellman equation, where the discriminator is introduced to assess differences between the predicted and actual rewards, guiding the learning process to focus on areas of high prediction error. By balancing estimation errors with a tuning parameter and carefully choosing model classes, the approach becomes robust against specific data patterns and high variance.

The IPW and DR methods can similarly be extended to Paradigms 2 and 3. For example, for IPW, we can replace the density ratio by that along the entire trajectory $\prod_{t'=0}^{t}[\pi(A_{i,t'}|S_{i,t'})/b(A_{i,t'}|S_{i,t'})]$. However, these traditional IS methods (and related DR methods) have exponential variance with the number of steps and hence will soon become unstable when the trajectory is long. To avoid this issue, various structural assumptions have been utilized. One notable advance is by considering the marginal importance ratio under stationary assumptions (Dai et al. 2020; Q. Liu et al. 2018; Shi, Wan, Chernozhukov, et al. 2021; J. Zhu et al. 2024), where essentially we consider the average density of visiting a state instead of considering the different densities at different time points, which allows us to greatly reduce the problem dimension. Similar techniques have been extended to the DR estimator as well (N. Jiang and L. Li 2016), notably the Double Reinforcement Learning method (Kallus and Uehara 2022)

7.2 Offline Policy Optimization

Another central task is to learn a good policy from the offline datasets. Formally, we want to find $\pi^* = \arg\max_\pi \eta(\pi)$. Similar to OPE, one key challenge is to adjust the selection bias caused by the distribution shift in offline datasets, therefore we will see similar tools in OPE (and also CEL) are extended here. We will additionally consider the unique goal of policy learning itself to design the so-called *pessimism*-based algorithms.

7.2.1 Model-Free Value-based Approaches. The first class of algorithms is *value*-based, focusing on first learning the value function and then optimizing the policy based on it. *Q-learning* is, arguably, the most popular algorithm due to its simplicity and good performance. Noting that $\pi^*(\boldsymbol{s}) = \arg\max_a Q(\boldsymbol{s}, a)$, the core of Q-learning is a regression modeling problem based on positing regression models for outcome. Overall, Q-learning is practical and easy to understand, as it allows straightforward implementation of diverse established regression methods. Different Q-function model classes (such as linear models, sparse linear models, neural networks, etc.) and their statistical properties have been studied extensively in the literature (R. Song, W. Wang, et al. 2015; W. Zhu et al. 2019).

In some cases, Q-learning could be overkill for policy optimization: for decision making, what truly matters is identifying the optimal action, which is analogous to knowing the expected potential outcomes of all actions. In such cases, Advantage-learning (A-learning) (Murphy 2003; Robins 2004; Schulte et al. 2014) offers a more efficient alternative by modeling only the contrasts between treatments and a control action, as $Q(\boldsymbol{s}, a) = Q(\boldsymbol{s}, 0) + A(\boldsymbol{s}, a)$. With the A-function $A(\boldsymbol{s}, a)$, we have that $\pi^*(\boldsymbol{s}) = \text{argmax}_{a\neq 0} A(\boldsymbol{s}, a)\mathbb{I}(\max_{a\neq 0} A(\boldsymbol{s}, a) > 0)$. Similar to Q-learning, various regression functions can be used to specify the advantage function. Typically, the underlying relationship in the advantage functions is simpler than that in $Q(\boldsymbol{s}, 0)$, which is a *nuisance* function in decision-making. The extension of A-learning to high-dimensional (Shi, Fan, et al. 2018) and non-parametric models (Liang et al. 2018) has also been studied.

Extensions to paradigms 2 and 3. These approaches can be similarly extended to MDP and non-Markovian decision processes. For example, the fitted-Q iteration (Ernst et al. 2005) extends the single-period Q-learning, by noting that the optimal value function Q^* is the unique solution to the Bellman optimality equation (1). Additionally, the right-hand side of (1) is a contraction mapping, allowing us to consider a fixed-point method similar to fitted-Q evaluation. A-learning is also recently extended to the MDP setting (Shi, S. Luo, et al. 2024).

Extending to non-Markovian problems, such as DTR in precision medicine and decision science, we can estimate the optimal dynamic treatment regimes (policy) via G-estimation (Robins, D. Lin, et al. 2004), a type of A-learning (Schulte et al. 2014; Shi, Fan, et al. 2018). Q-learning can also be extended to non-markovian problems via recursive regression (R. Song, W. Wang, et al. 2015). The main challenge comes from the increasing dimensionality with the expanding horizon, as the loss of the Markovian property requires us to use the full history in the feature space instead of only the latest state variable.

7.2.2 Model-Free Policy-based Approaches. The second class of algorithms is *policy*-based, which directly learns the policy without necessarily going through the learning of value functions. Specifically, for interpretability, domain constraints, or statistical efficiency, it may be preferable to directly learn a policy π within a pre-specified (parametric or non-parametric) policy class Π as $\pi^* = \arg\max_{\pi\in\Pi} \eta(\pi)$, where the policy value $\eta(\cdot)$ can be estimated via various OPE methods discussed above. This estimated value is then incorporated into an optimization process to solve the arg max via off-the-shelf *optimization algorithms* (such as the L-BFGS-B) (Kitagawa and Tetenov 2018; Y. Liu et al. 2018; Zhao, D. Zeng, et al. 2012).

In particular, when we use IPW as the policy value estimator, we can re-write the objective as

$$\pi^* = \arg\max_{\pi\in\Pi} \mathbb{E}\left[\frac{R_i}{b(A_i|S_i)}\mathbb{I}(A_i \neq \pi(S_i))\right].$$

When R_i is non-negative, this goal corresponds to the objective function of a cost-sensitive classification problem with $R_i/b(A_i|S_i)$ as the weight, A_i as the true label, and π as the classifier to be learned. Then, any popular cost-sensitive classifiers, such as support vector machine (SVM) and classification and regression trees (CART), can be applied to solve the policy learning problem. This is called outcome-weighted learning (Y. Liu et al. 2018; R. Song, M. Kosorok, et al. 2015; Zhao, D. Zeng, et al. 2012), providing flexibility in high-dimensional and complex scenarios. Furthermore, this framework can be extended to incorporate the DR estimator, enhancing robustness against misspecifications of the propensity score model or the outcome model.

Decision lists and tree-based structures are interpretable approaches in policy learning that provide a clear framework for treatment decisions within dynamic treatment regimes. Decision lists operate as sequential if-then rules, where each rule specifies conditions based on patient characteristics to guide treatment selection in a straightforward, deterministic manner (Tschernutter et al. 2022; Y. Zhang, Laber, et al. 2018). It is advantageous for transparency. For increased interpretability, sparse decision lists and pruned trees reduce model complexity, maintaining essential decision criteria without sacrificing clarity.

Extensions to paradigms 2 and 3. Extensions of this approach have also been developed to address more complex settings, such as paradigms 2 and 3, which involve multi-stage decision-making scenarios (P. Liao et al. 2022). Essentially, we replace the OPE methods in Paradigm 1 with their counterparts in Paradigms 2 and 3 introduced above.

7.2.3 Model-Based Approaches. Model-based reinforcement learning (MBRL) is a technique that leverages explicit models of the environment's dynamics to guide policy learning. This approach, rather than relying solely on observed rewards, uses a parameterized model to predict state transitions and rewards, enabling it to generate synthetic experiences that help train policies without direct interaction with the environment. Techniques in MBRL can include learning the dynamics of the environment to inform both planning and control, making it possible to learn policies even in complex, high-dimensional spaces (Deisenroth and Rasmussen 2011). These methods have shown effectiveness in offline reinforcement learning settings, as they mitigate the limitations of direct interaction by enabling supervised learning methods to fit the model and then use it for training or planning in a simulated environment (Levine, Finn, et al. 2016; Sutton 1991).

One key advantage of MBRL is its potential for sample efficiency, as it reuses past experiences by generating additional trajectories, thus enhancing policy learning. Additionally, model-based methods are versatile, as they

can integrate uncertainty estimation techniques to counteract distributional shifts. By estimating epistemic uncertainty, MBRL can prevent the exploitation of inaccuracies in the learned model. Recent studies, such as those using model-predictive control and policy rollouts, indicate promising results in high-dimensional tasks and show robust performance under various degrees of distributional shift, further affirming MBRL as a viable solution for offline policy learning (Chua et al. 2018; Nagabandi et al. 2018).

7.2.4 Address Increased Uncertainty from Distribution Shift. Besides the selection bias caused by the distribution shift in offline datasets, another prominent issue is the inflated uncertainty. The increased uncertainty results from the inherent limitations in the observational data used to inform policy decisions, which often fails to comprehensively represent the entire state and action space. Models trained on such non-representative datasets can yield overoptimistic predictions about the outcomes of actions, especially those that deviate substantially from the behavior policy used during data collection. As a result, suboptimal decisions would be made, as one policy may appear to be better just because its value estimate has a bigger variance. This problem is exacerbated in complex environments where the state and action spaces are vast and diverse, increasing the likelihood of encountering unrepresented scenarios.

To mitigate the risks associated with overoptimistic predictions, penalty-based or pessimism-based strategies are employed in offline policy learning. Penalty-based methods (Jaques et al. 2019) or constraint-based methods (Fujimoto et al. 2019; Kumar et al. 2019; Siegel et al. 2020) explicitly encourage or require the estimated optimal policy to stay close to the data distribution by introducing a penalty or constraint for taking actions that lead to high uncertainty. This penalty discourages the selection of such actions, steering the policy towards actions with more predictable outcomes based on the available data.

In contrast, pessimism-based methods (Jeunen and Goethals 2021; Rashidinejad et al. 2021; Y. Zhou et al. 2023) use an implicit and data-driven way to stay conservative and close to data distribution. It is typically based on explicit uncertainty quantification for the value estimates and then selects the policy that optimizes the value lower bounds. This approach thus reduces the likelihood of the algorithm recommending policies that just happen to be optimal due to high uncertainty. Theoretically, the pessimism-based algorithms can find an optimal policy when the data cover the trajectories of an optimal policy, an assumption that is much weaker than the full-coverage requirement.

In summary, the necessity for value pessimism, policy penalty, or policy constraint in offline policy learning arises from the need to counteract the inherent uncertainties associated with training models on limited observational datasets. By adopting these strategies, the reliability and safety of the policies derived from offline policy learning are enhanced, leading to more robust and effective decision making in practice.

8 Online Causal Policy Learning

This section explores strategies for addressing online decision-making problems under the data structures outlined in Paradigms 4-6, where the treatment policies are dynamically updated in real time based on data continuously collected through interactions with the environment. The primary goal in these settings is typically to optimize cumulative rewards. A key distinction between offline and online policy learning lies in the mode of data collection. While the performance of offline CPL could be constrained by the quality and representativeness of the already collected dataset, online CPL enables continual data collection, allowing policies to be progressively refined and adapted to nonstationary or evolving environments.

A central challenge in online CPL is the exploration-exploitation trade-off, which arises because only the outcomes corresponding to selected actions are observed, while outcomes under alternative actions remain unobserved. Addressing this challenge requires estimating the expected rewards of under-sampled or untried actions from partial, action-dependent data. This problem is inherently counterfactual in nature, as it involves reasoning about potential outcomes under hypothetical interventions. Consequently, even though many online

decision-making algorithms are not explicitly formulated within a causal inference framework, they fundamentally engage with causal issues related to missing counterfactuals, selective observation, and action-induced data bias.

In each of the following subsections of Section 8.1, we first briefly review classical online policy optimization methods that do not explicitly involve causal techniques. These methods fall under the direct method category, analogous to those commonly discussed in CEL, in which the reward or potential outcome is directly estimated as a function of actions (and possibly contexts), and decisions are made based on these estimates, implicitly assuming standard causal conditions like ignorability, positivity, and consistency. We then focus on the growing body of recent methods that explicitly leverage techniques inspired by causal inference to address the challenges inherent in online decision-making. In Section 8.2, we review existing approaches to online policy evaluation, which are predominantly formulated within causal inference frameworks.

8.1 Online Policy Optimization

We categorize online policy optimization problems into three groups based on their underlying causal structure assumptions. The first category encompasses problems that adhere to the data structure outlined in paradigm 4, and are widely studied as *bandit* problems (T. Lattimore and Szepesvári 2020; Slivkins et al. 2019), characterized by sequentially updated policies and independent state information. To address the complexities introduced by potential long-term dependencies between states, the second category considers the data structure of paradigm 5, assuming a Markovian state transition process, which has been extensively studied as *online RL* (Sutton and Barto 2018). The final category constitutes a broader class encompassing all remaining online learning problems characterized by non-Markovian system dynamics, as illustrated in paradigm 6, including those modeled by POMDPs (Meng et al. 2021; Spaan 2012) and DTR bandits (Y. Hu and Kallus 2020). POMDPs operate under the premise of an underlying MDP model, albeit with the challenge that the state itself is not directly observable. Conversely, DTR bandits leverage the entire history of past information to iteratively learn an optimal treatment regime, typically within a short horizon due to computational complexity. While traditional research on these directions rarely frames its methodologies in causal terms, it implicitly relies on counterfactual reasoning. Recently, growing recognition of this causal nature has led to increasing efforts to integrate classical causal inference techniques to enhance the robustness, generalizability, and interpretability of online decision-making (Deng et al. 2023; Y. Zeng et al. 2024).

Table 5. Summary of Online CPL Literature (Paradigm 4 without Explicit Use of CEL/CSL Techniques)

Method Type	Application	Paper
Value-Based	MAB	ϵ-greedy (Auer et al. 2002; Sutton and Barto 1999); UCB (Auer et al. 2002); TS (Agrawal and Goyal 2013a; Chapelle and L. Li 2011; Wan, Wei, et al. 2023)
	Contextual MAB	UCB (Chu et al. 2011; L. Li, Chu, Langford, and Schapire 2010); TS (Agrawal and Goyal 2013b; Kveton, Zaheer, et al. 2020)
	Multi-Task MAB	meta-TS (Kveton, Konobeev, et al. 2021; Wan, Ge, et al. 2021)
	Structured Bandits	UCB (W. Chen et al. 2013; Kveton, Szepesvari, et al. 2015); TS(Agrawal, Avadhanula, et al. 2017); meta-TS (Wan, Ge, et al. 2023)
Policy-Based	MAB	Gradient Bandit Algorithms(Mei et al. 2023; Sutton and Barto 2018)
	Contextual MAB	Classification Oracle-Based Algorithms (Agarwal, Hsu, et al. 2014; Dudik et al. 2011; Langford and T. Zhang 2007)

8.1.1 Bandits (Paradigm 4). Bandit problems have been widely used in a variety of fields, including recommender systems (Q. Zhou et al. 2017), clinical trials (Durand et al. 2018), and business or economic applications (W. Shen et al. 2015). In a bandit problem, an agent sequentially selects actions A_t from a (possibly time-varying) action set $\mathcal{A}$ and observes corresponding rewards R_t. The ultimate objective is to maximize the cumulative reward $\sum_{t=0}^{T} R_t(A_t)$, or equivalently, minimize the cumulative regret $\sum_{t=0}^{T} \max_{a\in\mathcal{A}} R_t(a) - R_t(A_t)$. A fundamental challenge in this setting is the uncertainty of the counterfactual reward distributions, which can only be learned through sequential interaction with the environment. At the heart of it lies the exploration–exploitation trade-off.

The simplest and most studied setting is the MAB problem (Bouneffouf et al. 2020; Slivkins et al. 2019), where the action space consists of a finite set of K actions, or "arms" (i.e., $\mathcal{A} = \{1, 2, \ldots, K\}$). Classical algorithms for MAB can be broadly grouped into three categories. **I) ϵ-greedy algorithms** (Auer et al. 2002; Sutton and Barto 1999) alternate between exploiting the empirically best arm and random exploration, offering simplicity but limited statistical efficiency due to their inability to explicitly model uncertainty. **II) Upper Confidence Bound (UCB)-type algorithms** (Auer et al. 2002; Slivkins et al. 2019) adopt a frequentist perspective and embody the principle of *optimism in the face of uncertainty*. At each round, an arm is selected according to its upper confidence bound, defined as the sum of the estimated mean reward and a confidence radius that quantifies uncertainty. This strategy achieves near-optimal regret guarantees. **III) Thompson Sampling (TS)** (Agrawal and Goyal 2013a) follows a Bayesian approach by sampling rewards from posterior distributions over arms, naturally balancing exploration and exploitation. It often performs favorably in practice but can be sensitive to prior misspecification (Chapelle and L. Li 2011; T. Lattimore and Szepesvári 2020; Wan, Ge, et al. 2023).

Beyond the basic MAB setting, numerous extensions have been developed to address real-world complexities. Contextual bandits incorporate side information to personalize decisions (Agrawal and Goyal 2013b; Chu et al. 2011; L. Li, Chu, Langford, and Schapire 2010); structured bandits exploit known dependencies or combinatorial structures in the action space (Agrawal, Avadhanula, et al. 2017; W. Chen et al. 2013; Kveton, Szepesvari, et al. 2015); and multi-task bandits enable information sharing across related decision problems to accelerate learning (Kveton, Konobeev, et al. 2021; Wan, Ge, et al. 2021).

All of the methods above are **value-based**, in that they estimate action rewards and select actions accordingly. In contrast, **policy-based** approaches learn action preferences or policies directly. For example, gradient bandit methods (Mei et al. 2023; Sutton and Barto 2018) maintain preference scores (or functions) over actions. In contextual settings, classification-oracle-based methods (Agarwal, Hsu, et al. 2014; Dudik et al. 2011; Langford and T. Zhang 2007) directly optimize policies over a predefined policy class by reducing policy learning to cost-sensitive multi-class classification problem.

Bandits with Causality. Recent advances in CSL and CEL offer a promising opportunity to increase the learning efficiency of bandits. For example, CEL methods like IPW and DR estimation have been adapted for value estimation to mitigate potential biases in bandit learning resulting from reward model misspecification or covariate imbalances—particularly when training data is sparse or unrepresentative in specific areas of the context space. Two primary approaches have emerged for integrating IPW/DR into bandit algorithms: (i) generating unbiased pseudo-rewards from observed rewards and propensity scores, adhering to the principles of IPW/DR estimators in CEL, which are subsequently used in the bandit update process in place of the observed rewards (Bietti et al. 2021; G.-S. Kim and Paik 2019; W. Kim, G.-S. Kim, et al. 2021; W. Kim, K. Lee, et al. 2023); and (ii) employing importance-weighted regression, wherein each observation is weighted by the inverse of its propensity score (Bietti et al. 2021; Dimakopoulou, Z. Zhou, et al. 2019). While these methods remain focused on maximizing rewards, recent research has also explored integrating classical bandit frameworks with meta-learners to optimize the incremental benefit of an action (e.g., $\tau_a = R(A = a) - R(A \neq a)$), aiming to enhance the return on investment (Kanase et al. 2022; Zhao, Goodman, et al. 2022) or potentially simplify model estimation (Carranza et al. 2023).

Table 6. Summary of Online CPL Literature (Paradigm 4 with Explicit Use of CEL/CSL Techniques)

Application	Method Type	Paper
CEL/CSL for Bandit Optimization	Reward Model Debiasing	Bietti et al. (2021), Carranza et al. (2023), Dimakopoulou, Z. Zhou, et al. (2019), Kanase et al. (2022), G.-S. Kim and Paik (2019), W. Kim, G.-S. Kim, et al. (2021), W. Kim, K. Lee, et al. (2023), and Zhao, Goodman, et al. (2022)
	Knowledge Transfer	Y. Li et al. (2021), H. Xu and Xie (2023), and J. Zhang and Bareinboim (2017)
	Exploiting Causal Structures in Multi-Intervention Settings	F. Lattimore et al. (2016), S. Lee and Bareinboim (2019, 2018), Y. Lu et al. (2020), Nair et al. (2021), Sen, Shanmugam, Dimakis, et al. (2017), and Subramanian and Ravindran (2021)
CEL/CSL for Bandit Evaluation (See detailed discussions in Section 8.2)	-	Bibaut et al. (2021), Chambaz et al. (2017), H. Chen et al. (2020), Dimakopoulou, Ren, et al. (2021), Khamaru et al. (2021), Ramprasad et al. (2023), Y. Shen et al. (2024), Zhan et al. (2021), and K. Zhang et al. (2021)

Another line of recent research leverages causal techniques to transfer knowledge from logged data to "warm up" bandit agents. This approach initiates agents with an informative estimate of the environment, thereby reducing the number of rounds required for online exploration. Y. Li et al. (2021) propose creating a pseudo-environment using logged data to synthesize action outcomes via matching and weighting and introduce a two-stage learning process under the UCB framework. Specifically, in the first stage, the pseudo-environment is used to simulate interactions with the bandit agent in order to prepare the agent for real-world engagement. In the second stage, the bandit agent uses the knowledge gained in the first stage to interact with the real world, significantly reducing the possibilities of unnecessary exploration. Similarly, H. Xu and Xie (2023) provides a complementary view with a TS-inspired variant. Distinct from creating a pseudo-environment, J. Zhang and Bareinboim (2017) employ SCMs to derive causal bounds for potential outcomes, facilitating the transfer of learnings between bandit agents. By leveraging the causal structure of the environment and the observed trajectories from completed bandit agents, they propose to derive the two-sided bounds on the potential rewards over the action space for the target bandit agent. These bounds are subsequently utilized to eliminate less effective options during the initialization phase and to refine the UCB estimates throughout the learning process.

Furthermore, side causal information is particularly effective in improving learning efficiency in scenarios with multiple intervention variables. For example, F. Lattimore et al. (2016) is the first to introduce such a class of causal bandit problems. Given a causal graph among variables that include either interventional or non-interventional variables and reward, agents are able to select more than one variable in the causal graph to intervene at each round. Utilizing the causal graph to transfer information among interventional variables and hence reduce the amount of exploration needed, F. Lattimore et al. (2016) and Sen, Shanmugam, Dimakis, et al. (2017) focus on the best arm identification problem, while Y. Lu et al. (2020) and Nair et al. (2021) propose algorithms to minimize the cumulative regret. However, when considering a large number of interventional variables, S. Lee and Bareinboim (2018) empirically showed that a brute-force way to apply standard bandit algorithms to all interventions can suffer high regret. To further enhance sampling efficiency, S. Lee and Bareinboim (2019, 2018) proposed narrowing the action space by determining the possibly optimal minimal intervention set before applying standard bandit algorithms, while Subramanian and Ravindran (2021) suggested performing target interventions to allocate more samples to targeted subpopulations most informative about the most valuable interventions.

Beyond improving learning efficiency, causality also addresses broader challenges in bandit learning, including robustness to assumption violations (Section 9) as well as fairness and explainable decision making (Section 11).

Table 7. Summary of Online CPL Literature (Paradigms 5 & 6)

Method Type	Represented Paper
Value-Based	Monte Carlo Sampling(S. P. Singh and Sutton 1996); SARSA (Rummery and Niranjan 1994); Q-learning (Watkins and Dayan 1992); A-learning (Baird 1994)
Policy-Based	Schulman, Levine, et al. (2015), Schulman, Wolski, et al. (2017), and Williams (1992)
Actor-Critic	Gu, Lillicrap, et al. (2017), Mnih, Badia, et al. (2016), and Schulman, Moritz, et al. (2015)
Efficient Exploration via CEL/CSL techniques	X. Hu et al. (2022) and Seitzer et al. (2021)

8.1.2 Online Reinforcement Learning (Paradigms 5 & 6). Unlike bandit problems, which assume actions have only immediate effects, online RL explicitly models long-term consequences through state transitions (Sutton and Barto 2018), making it central to applications such as robotics and autonomous driving (Kiran et al. 2021; B. Singh et al. 2022).

Similar to offline RL, online RL encompasses value-based and policy-based approaches. **Value-based methods** learn Q/V functions and derive policies accordingly. Classical examples include Monte Carlo sampling (S. P. Singh and Sutton 1996) and temporal-difference learning (Sutton 1988), with representative algorithms such as SARSA (Rummery and Niranjan 1994) and Q-learning (Watkins and Dayan 1992). **Policy-based methods** directly optimize parameterized policies by maximizing expected cumulative reward, as in REINFORCE, PPO, and TRPO (Schulman, Levine, et al. 2015; Schulman, Wolski, et al. 2017; Williams 1992), offering flexibility and natural extensions to POMDPs (Paradigm 6), but often suffering from high variance and sample inefficiency. To address these limitations, **actor–critic algorithms** combine value-based and policy-based approaches by maintaining both a policy estimator (actor) and a value estimator (critic), using value estimates to guide policy updates (Gu, Lillicrap, et al. 2017; Mnih, Badia, et al. 2016; Schulman, Moritz, et al. 2015). While these methods are closely related to offline RL, the defining challenge in online settings is exploration: agents must actively collect interaction data to improve policies and mitigate distributional shift, commonly using ϵ-greedy (Mnih, Kavukcuoglu, et al. 2015), UCB (M. Bellemare et al. 2016), and TS (Osband et al. 2016).

Online RL with Causality. More recently, causal graph structures have been incorporated into online RL to guide exploration in a more principled and data-efficient manner. Rather than relying on undirected randomness or generic uncertainty estimates, causal approaches explicitly reason about how agent actions influence environment dynamics. For example, Seitzer et al. (2021) introduces a framework that employs situation-dependent causal influence, measured via conditional mutual information, to identify states where an agent can effectively influence its environment. Integrating this measure into RL algorithms enhances exploration and off-policy learning, significantly improving data efficiency in robotic manipulation tasks. Similarly, X. Hu et al. (2022) proposes the Causality-Driven Hierarchical RL framework, which leverages causal relationships among environment variables to discover and construct high-quality hierarchical structures for exploration, thereby avoiding inefficient randomness-driven exploration.

Although the use of causal methods in online RL specifically for improving sampling efficiency remains relatively limited, the integration of causal reasoning and online RL is an active research area spanning multiple directions. For example, causal modeling has also been explored in the context of handling assumption violations (Section 9) and promoting fairness and explainability in sequential decision making (Section 11).

8.2 Online Policy Evaluation (Paradigm 4 & 5)

The evaluation of the performance of policies plays a vital role in many areas, including medicine and economics (see e.g., Athey 2019; Chakraborty and Moodie 2013). By evaluation, we aim to unbiasedly estimate the value of the optimal policy that the bandit policy is approaching and infer the corresponding value. Although there is an increasing research interest in policy evaluation (see e.g., Dudík et al. 2011; N. Jiang and L. Li 2016; Kallus and A. Zhou 2018; L. Li, Chu, Langford, and X. Wang 2011; Y. Su et al. 2020; Swaminathan et al. 2017; Y.-X. Wang et al. 2017), we note that all of these works focus on learning the value of a target policy offline using historical log data. Instead of post-experiment investigation, increasing attention has recently been given to evaluating the ongoing policy in real time.

Despite the importance of policy evaluation in online learning, current bandit literature suffers from three main challenges. First, the data, such as the actions and rewards sequentially collected from the online environment, are not independent and identically distributed (i.i.d.), as they depend on the previous history and the running policy. In contrast, existing methods for offline policy evaluation (see e.g., Dudík et al. 2011; L. Li, Chu, Langford, and X. Wang 2011) primarily assume that the data are generated by the same behavior policy and are i.i.d. across individuals and time points. The second challenge lies in estimating the mean outcome under the optimal policy online. Although numerous methods have recently been proposed to evaluate the online sample mean for a fixed action (see e.g., Deshpande et al. 2018; Hadad et al. 2021; Neel and A. Roth 2018; Nie, Tian, et al. 2018; Shin et al. 2019, 2021; K. Zhang et al. 2020), none of these methods are directly applicable to online policy evaluation, as the sample mean only provides the impact of one particular arm, not the value of the optimal policy in bandits that consider the dynamics of the online environment. Third, given data generated by an online algorithm that maintains the exploration-exploitation trade-off sequentially, inferring the value of a policy online should account for this trade-off and quantify the probabilities of exploration and exploitation.

There are very few studies directly related to the topic of online policy evaluation. Chambaz et al. (2017) established the asymptotic normality for the conditional mean outcome under an optimal policy for sequential decision making. Later, H. Chen et al. (2020) proposed an inverse probability weighted value estimator to infer the value of the optimal policy using the ϵ-Greedy method. Recently, to evaluate the value of a known policy based on adaptive data, Bibaut et al. (2021) and Zhan et al. (2021) proposed to utilize the stabilized doubly robust estimator and the adaptive weighting doubly robust estimator, respectively. Both methods focused on obtaining valid inference for the value estimator under a fixed policy by conveniently assuming a desired exploration rate to ensure sufficient sampling of different arms. Also see other recent advances that focus on statistical inference for adaptively collected data (Dimakopoulou, Ren, et al. 2021; Ramprasad et al. 2023; K. Zhang et al. 2021) in bandit or RL settings. To infer the value of the optimal policy by investigating the exploration rate in online learning, Y. Shen et al. (2024) explicitly characterizes the trade-off between exploration and exploitation in online policy optimization, by deriving the probability of exploration in bandit algorithms. Their work proposed the doubly robust interval estimation (DREAM) method to infer the mean outcome of the optimal online policy with double protection. We note that these existing approaches for online policy evaluation are predominantly formulated within causal inference frameworks with data collected from online environment.

9 Scenarios Violating Causal Assumptions

In the sections above, most of the literature proceeds under the causal identification assumptions outlined in Section 3.4. However, real-world scenarios sometimes violate these assumptions. For instance, the SUTVA assumption assumes that a unit's outcome isn't influenced by the treatment of other units. Yet, for example, in the spread of COVID-19, if we simply assume each individual as one unit, the SUTVA is violated as each patient's health can be affected by the immunization status of the entire community; the NUC assumption is commonly violated in observational data due to those unmeasured confounders that researchers cannot fully

control without experimental design; the positivity assumption is violated when some individuals can not receive a specific treatment or control with certainty due to ethical issues or budget constraints.

It is crucial to handle these cases for better decision making, as accurately estimating causal effects in the face of assumption violations is usually the first step towards optimizing policies for higher rewards. Recently, the literature has started to focus on addressing each assumption violation using various tools, most of which were originally developed in the area of causal inference. In this section, we will provide a concise overview and outline open questions in the literature that need further exploration.

9.1 Unmeasured Confounders

Table 8. A brief summary of papers where unmeasured confounders exist

Paradigm	Instrumental Variables	Data Integration
Paradigm 1	Angrist, G. W. Imbens, and Rubin (1996) L. Wang and Tchetgen Tchetgen (2018) Qiu, Carone, Sadikova, et al. (2021) Cui and Tchetgen Tchetgen (2021)	Wu and S. Yang (2022)
Paradigm 2/5	L. Liao et al. (2024) and Y. Xu, J. Zhu, et al. (2023)	G. Imbens et al. (2024)
Paradigm 3/6	S. Chen and B. Zhang (2023) Y. Xu, J. Zhu, et al. (2023)	Athey, Chetty, et al. (2020)
Paradigm 4	Kallus (2018)	Bareinboim et al. (2015) Sen, Shanmugam, Kocaoglu, et al. (2017) L. Xu et al. (2021)

Table 8 provides a brief summary of some representative approaches in handling the violation of the NUC assumption. When there is an unmeasured confounder U, traditional methods exist for CEL and CPL would result in biased estimates. To adjust for potential bias, the current literature can be broadly divided into two categories: (1) using proxies such as Instrumental Variables (IV), or (2) incorporating additional data sources, typically integrating experimental data without unmeasured confounders with existing observational data (commonly referred to as data integration).

The use of IV can be traced back to the 1920s (Wright 1928), gaining widespread recognition with the introduction of the Two-Stage Least Squares (2SLS) method (Angrist and G. W. Imbens 1995). Typically, a variable X is called an IV when it satisfies the following three conditions:

a. IV independence: $Z \perp U|S$.
b. IV relevance: $Z \not\perp A|S$.
c. Exclusion restriction: $R(z, a) = R(a)$ for any $(z, a) \in \mathcal{Z} \otimes \mathcal{A}$.

In particular, IV has been used without stringent modeling assumptions to effectively estimate ATEs (Angrist, G. W. Imbens, and Rubin 1996), where a certain monotonicity assumption is imposed to guarantee identifiability. In a single-stage setting (Paradigm 1), recent advancements, such as L. Wang and Tchetgen Tchetgen (2018), have focused on developing semi-parametric efficiency estimators for HTE with IV. Additionally, Qiu, Carone, Sadikova, et al. (2021) and Cui and Tchetgen Tchetgen (2021) have contributed to the literature by deriving optimal policies that maximize rewards in the presence of IV.

Under MDPs (Paradigms 2 or 5), the existence of unobserved state variables also greatly influences both the estimation and decision-making process. This problem has been explored in the literature of RL, as evidenced by works such as Y. Xu, J. Zhu, et al. (2023) and G. Imbens et al. (2024). In offline RL, Y. Xu, J. Zhu, et al.

(2023) introduced IV-based methods to ensure consistent OPE in confounded sequential decision making, which emphasizes semi-parametric efficiency, statistical inference, and extensions to high-order MDPs and POMDPs. In addition to IVs, another line of research aims to combine interventional data and observational data for better decision making. For example, G. Imbens et al. (2024) combined short-term experimental data and long-term observational data with potential confounders to handle the identification and estimation of long-term treatment effects with guarantees from asymptotic theory.

In scenarios where the Markov assumption does not hold (Paradigm 3), addressing confounders requires tailored approaches depending on differences in data-generating structures. For example, Shi, Uehara, et al. (2022) proposed an approach OPE within the framework of POMDPs, which is a natural extension of MDPs to handle unmeasured confounders present in the latent state. In DTR, S. Chen and B. Zhang (2023) also employed IV for consistent treatment effect estimation and policy optimization. When only a short-term variable (i.e., a surrogate) is observed for long-term treatment effect estimation, and there are no clear state transitions defined as DTR or MDPs, the data structure introduced by Athey, Chetty, et al. (2020) is more suitable for identifying and estimating long-term effects using short-term experimental data.

In online bandits learning (Paradigm 4), some work has also been done with IV to help detect possible unmeasured confounders and avoid biased policy learning (Bareinboim et al. 2015; Kallus 2018; Sen, Shanmugam, Kocaoglu, et al. 2017; L. Xu et al. 2021). In the context of causal inference, the bandits problem is also equivalent to first estimating $R(a)$ and adding appropriate exploration to obtain a suboptimal regret bound. When there are confounders, at least part of the relationship between A and R is not captured in the reward modeling process, leading to biased reward modeling and, consequently, biased policy learning outcomes. Given the flexible approaches in handling unmeasured confounders in CEL, recent research has been developed to address this problem. Most of the existing literature focuses on introducing proxy variables, such as IVs (Kallus 2018), where authors investigate optimal policies to maximize intent-to-treat regret in the presence of potential non-compliance and unmeasured confounders, or combining observational data for confounding adjustment within the structural causal equation model (Bareinboim et al. 2015). Additionally, L. Xu et al. (2021) proposed a two-stage regression scheme based on proxy variables to handle unmeasured confounders, especially when the data are high-dimensional and non-linear.

9.2 Interference

Table 9. A brief summary of papers when interference exists

Paradigm	Methogologies
Paradigm 1	Experimental design: Aronow and Samii (2017), Leung (2022a,b), and Viviano (2024) Observational studies: Bargagli-Stoffi et al. (2025), Forastiere et al. (2021), and L. Su et al. (2019)
Paradigm 2/5	S. Luo et al. (2024), Shi, Wan, G. Song, et al. (2023), and Y. Yang, R. Luo, et al. (2018) Multi-agent RL: M. Chen et al. (2021) and Y. Yang, Ma, et al. (2021) J. Jiang and Z. Lu (2023) and Pan et al. (2022)
Paradigm 3/6	DTR with interference: C. Jiang et al. (2023)
Paradigm 4	Bargiacchi et al. (2018) and Verstraeten et al. (2020) Multi-agent bandits: Agarwal, Agarwal, et al. (2024) and Dubey et al. (2020) Jia et al. (2024) and Y. Xu, W. Lu, et al. (2024)

Table 9 provides a brief summary of some representative approaches in handling the violation of the SUTVA assumption. Interference, often known as the existence of spillover effect (SE), is a commonly encountered problem in causal inference. Generally, it requires extending the SUTVA assumption in that the potential outcomes of unit

i depend on not only A_i, but also the actions of other units. For example, the reward of unit i under interference is denoted by $R_i(\boldsymbol{A})$, where $\boldsymbol{A} = \{A_1, \ldots, A_n\}$. Following this definition, the treatment effect can be decomposed into two parts: DE and SE :

$$\mathrm{DE}_i(a_i, a_i', \boldsymbol{a}_{-i}) = R_i(a_i, \boldsymbol{a}_{-i}) - R_i(a_i', \boldsymbol{a}_{-i}),$$
$$\mathrm{SE}_i(a_i, \boldsymbol{a}_{-i}, \boldsymbol{a}_{-i}') = R_i(a_i, \boldsymbol{a}_{-i}) - R_i(a_i, \boldsymbol{a}_{-i}'),$$

where $\boldsymbol{a}_{-i}$ denotes the action assignment vector for all units except unit i.

Due to this dependency, directly modeling the treatment effect without considering the actions of other units can lead to bias. However, in extreme situations where each unit's reward depends on every single unit's action, any reward modeling approach to generalize findings across units would fail, making causal effect identification impossible. Consequently, there is a growing trend in the literature toward identifying and estimating direct effect (DE) and spillover effect (SE) under various model structures, aiming to strike a balance between avoiding overly stringent assumptions on interference structure and allowing learning from existing data.

In the existing literature, various interference structures have been considered, including, but not limited to, partial interference (Sobel 2006), stratified interference (Hudgens and Halloran 2008), neighborhood interference (Forastiere et al. 2021), spatial interference (Leung 2022b), and cluster network interference (Bargagli-Stoffi et al. 2025). Despite various definitions, most types of interference share a similar two-step structure to simplify the problem. First, units are typically categorized into groups through clustering or partitioning, assuming that interference occurs only within each group. Extensions of this assumption allow for interference between any units, where the degree of interference decreases with distance but is not necessarily zero (Leung 2022a,b). Second, to further simplify causal identification and estimation within each cluster, where the strength of interference level may vary by domain assumption, interference is also commonly quantified by *exposure mapping*. This concept, proposed by Aronow and Samii (2017), is similar to the "effective treatments" function in Manski (2013). Generally, exposure mapping assumes that interference among units is passed through a lower-dimensional exposure mapping function, often known and assumed to follow a parametric form to easily quantify the SE between interfering units. This approach has been widely adopted in various papers under specific mapping forms, such as L. Su et al. (2019) and Viviano (2024).

Depending on the quantity of interest and the flexibility of managing treatment assignment, different methods have been developed to address various application needs. There is a growing trend of research that focuses on estimating causal DEs and SEs in both (1) experimental design (Aronow and Samii 2017; Leung 2022a,b; Viviano 2024) and (2) observational studies (Bargagli-Stoffi et al. 2025; Forastiere et al. 2021; L. Su et al. 2019). Specifically, among the approaches focusing on (1) experimental design, Leung (2022b) concentrates on developing optimal designs to maximize specific goals, such as achieving near-optimal rates of convergence for global average causal effect estimation or minimizing the variance of the estimators. In contrast, Aronow and Samii (2017), Leung (2022a), and Viviano (2024) focus more on estimation under randomized experiments, providing asymptotic guarantees for the proposed estimators for both average effects (Aronow and Samii 2017; Leung 2022a) and heterogeneous effects (Viviano 2024). Among the approaches focusing on (2) observational studies, methods vary based on assumptions about the type of interference and reward modeling. For example, L. Su et al. (2019) considers the reward as a linear function of neighbors' covariates and treatments, extending Q-learning and A-learning to scenarios with interference under certain structural assumptions. Forastiere et al. (2021) and Bargagli-Stoffi et al. (2025) emphasize semi-parametric or non-parametric estimation and inference under looser modeling assumptions. Users may select the appropriate method according to the assumptions they are willing to make regarding interference structures and modeling strategies.

Moving to multi-stage settings, including MDPs (Y. Yang, R. Luo, et al. 2018), DTR (C. Jiang et al. 2023) and beyond, interference often arises in a multi-agent system. In Multi-Agent RL (MARL), the concept of neighborhood interference and exposure mapping is incorporated into the Q function. Here, the Q value of agent j depends not

only on (s_j, a_j) but also on the actions of a neighborhood set $\boldsymbol{a}_{\mathcal{N}_j}$, which is the so-called mean-field approximation strategy (Y. Yang, R. Luo, et al. 2018). Later on, this mean-field approximation method has been applied in various MARL studies, including M. Chen et al. (2021), S. Luo et al. (2024), and Shi, Wan, G. Song, et al. (2023). Among these, Shi, Wan, G. Song, et al. (2023) and S. Luo et al. (2024) focus more on OPE with observational data in the presence of both temporal and/or spatial dependencies among agents. Other related work in offline MARL includes, but is not limited to Y. Yang, Ma, et al. (2021), Pan et al. (2022), and J. Jiang and Z. Lu (2023). Specifically, Pan et al. (2022) tackles challenges posed by an increasing number of agents on conservative offline RL algorithms, J. Jiang and Z. Lu (2023) exploits value deviation and transition normalization in non-stationary transition dynamics, and Y. Yang, Ma, et al. (2021) focuses on mitigating extrapolation error in offline evaluation.

The multi-agent system is also present in online bandits (paradigm 4) (Bargiacchi et al. 2018; Dubey et al. 2020; Verstraeten et al. 2020). In round t, each agent i is able to interact with the environment and pull an arm. Interference exists since the actions of agents are mutually affected by each other due to the neighboring relationships among agents. Similar to single-stage settings, a common approach to handle this problem is to decompose the agents into fixed but potentially overlapping subgroups. This decomposition simplifies the reward of the joint action space to the summation of local reward functions, reducing the complexity of global exploration. Following this approach, Bargiacchi et al. (2018) extended the UCB algorithm from classical MAB to the multi-agent scenario. Shortly after, Verstraeten et al. (2020) extended the TS algorithm to the multi-agent case under similar interference assumptions. Dubey et al. (2020) considered another reward modeling structure where interference is transmitted through network contexts, and proposed a kernelized UCB algorithm to cooperatively maximize cumulative rewards. Rather than handling interference implicitly within multi-agent systems, recent pioneering works have explicitly addressed the interference issue in bandit settings. For example, Agarwal, Agarwal, et al. (2024) studies a sparse network interference model under the multi-armed bandits framework, while Y. Xu, W. Lu, et al. (2024) examines interference in contextual bandits. Despite these advancements, this field remains relatively underexplored, highlighting a pressing need for more flexible and general approaches to address broader bandit settings using causal methodologies.

9.3 Positivity

The positivity (or overlap) assumption is a fundamental requirement to ensure that the regions of the data space covered by the target policy are adequately represented in the observed data. This assumption posits that the conditional probability that any unit group will take each action is strictly greater than zero. However, in observational studies where actions are not controlled a priori, satisfying this assumption is often challenging due to several factors: (1) in continuous action spaces, it is naturally impossible for an observational study to exhaustively cover all actions; (2) when the feature space for state information $\boldsymbol{S}$ is high-dimensional, ensuring sufficient overlap becomes difficult; and (3) specific treatment options might lack observational data due to ethical concerns, high costs, or oversights during data collection. These issues are prevalent in offline studies (paradigms 1-3), whereas in online experiments, actions are typically selected by designers, thereby mitigating the non-overlapping problem.

Although the violation of the positivity assumption receives less attention compared to the previous two assumptions, existing literature still offers some solutions to address this issue. Given that the general strategies are quite similar across different data structures and that this problem is generally less studied in paradigms 2 and 3, we will focus on introducing approaches to handle the violation of the positivity assumption under paradigm 1.

In the three scenarios mentioned, continuous action space and high-dimensional covariate space are cases where existing papers may not specifically address non-overlapping issues. However, these approaches provide a general framework that naturally tackles this problem. For example, D'Amour et al. (2021) discuss the trade-off between incorporating more covariates to mitigate unmeasured confounders and the difficulty of satisfying the

positivity assumption. They argue that strict overlap is more restrictive than expected in many studies and derive explicit bounds on the average imbalance in covariate means.

In addressing continuous action space, most existing work tackles this problem through non-parametric smoothing (G. Chen et al. 2016; Kallus and A. Zhou 2018) or by combining it with (semi)-parametric modeling (Chernozhukov, Demirer, et al. 2019). For instance, in the IPW estimator, Kallus and A. Zhou (2018) handle both policy evaluation and optimization by smoothly relaxing the indicator function using a kernel function, while G. Chen et al. (2016) replace the indicator function with a hinge loss. The continuous action space, as a well-established research area in CDM, has extensive literature beyond the examples listed, and a more detailed review is beyond the scope of this paper.

The last stream of work directly addresses non-overlapping issues without assuming specific dimensions for the action or covariate space. Due to the scarcity of data points in low-overlap areas, inferring causal relationships in these regions is challenging without additional smoothing or parametric modeling assumptions. Traditional approaches often use trimming (Crump et al. 2009; Petersen et al. 2012; Rosenbaum and Rubin 1983), which involves discarding units with estimated propensity scores outside a specific range $[0.1, 0.9]$, or matching (Visconti and Zubizarreta 2018). Recently, in OPE, Khan et al. (2024) applied partial identification techniques from causal inference to derive OPE results under non-parametric Lipschitz smoothness assumptions on the reward function. In OPO, the pessimism technique becomes particularly important due to the increased uncertainty for candidate policies with poor coverage. For example, Jin et al. (2023) used pessimism with the generalized empirical Bernstein's inequality to study OPO without the uniform overlap condition.

10 Real Data

In this section, we first present two representative examples, the Medical Information Mart for Intensive Care III (MIMIC-III)[2] and the MovieLens[3] datasets, to illustrate the tasks and paradigms discussed in the previous sections. These examples offer concrete demonstrations of how the complete CDM process is applied in real-world contexts, such as clinical trials and recommendation systems. In Section 10.3, we summarize commonly used software packages that support various tasks in causal decision making.

It is important to note that the goal of this section is not to compare the performance of different methods for each task but rather to demonstrate the integration of all components within a unified causal decision making framework. Our intention is to tell a cohesive story that walks through the entire structure of causal decision making, while highlighting methods and software packages that are readily accessible to practitioners for further exploration.

10.1 MIMIC-III

The MIMIC-III dataset (Johnson et al. 2016) is a large, de-identified, and publicly available collection of medical records that contains comprehensive clinical data from patients admitted to the intensive care unit (ICU) at a large tertiary care hospital. Of particular interest are the records of patients with sepsis, a life-threatening response to infection caused by harmful microorganisms, which significantly contribute to clinical research and practice. This disease accounts for over 200,000 annual deaths in the U.S. alone, underscoring the urgent need for effective and timely interventions. Given its treatable nature, sepsis demands prompt emergency care and robust decision-making frameworks to improve patient outcomes and reduce mortality.

However, the electronic health record (EHR) collected from patients with sepsis in ICUs presents significant challenges to the application of existing decision-making methods. Specifically, records of hundreds of thousands of sepsis patients treated at the Beth Israel Deaconess Medical Center between 2001 and 2012 include numerous

[2] https://www.kaggle.com/datasets/asjad99/mimiciii
[3] https://grouplens.org/datasets/movielens/1m/

covariates such as demographics, vital signs, medical interventions, lab test results, and post-treatment outcomes. For researchers, it is essential to disentangle the complex causal relationships among these variables, understand the impacts of specific sepsis treatments, and select a necessary and sufficient set of variables to analyze the disease. This comprehensive causal understanding is critical to optimizing treatments and ultimately reducing mortality associated with sepsis.

To overcome these challenges, we utilize the three main tasks mentioned above and demonstrate the causal decision-making process using MIMIC-III as an example. The first step, CSL, which aims to uncover the causal relationships between variables, allows us to pinpoint the right and informative set of state variables, the treatments, and mediators that may influence the mortality due to sepsis. In this context, we identify the IV-Input as an actionable treatment, while the Sepsis-related Organ Failure Assessment (SOFA) score serves as an important mediator, describing organ dysfunction or failure. The second step, CEL, building upon the causal relationships identified in CSL, quantifies the strength of causal links and thus measures the most effective treatments and informative mediators. For example, it estimates the average treatment effect to determine how intravenous fluid input (IV-Input) impacts mortality rates across the general patient population and the heterogeneous treatment effect for individual patients. The final step, CPL, seeks to determine the optimal administration policy to minimize patient mortality rates. Each patient visit is treated as a stage, and depending on the MIMIC-III data structure, various CPL algorithms can be applied to the appropriate paradigm to learn the optimal policy from observational data.

Due to privacy concerns, we used a subset of the original data, which is publicly available on Kaggle. For illustration purposes, we selected several representative features for the following analysis, as listed in Table 10.

Table 10. Description of the variables used in the MIMIC-III data analysis

Variable	**Description**
Glucose	Glucose values of patients
PaO2_FiO2	The partial pressure of oxygen (PaO2) divided by the fraction of oxygen delivered (FiO2)
SOFA	Sepsis-related Organ Failure Assessment score to describe organ dysfunction/failure
IV-Input	The volume of fluids that have been administered to the patient
Died_within_48H	The mortality status of the patient 48 hours after treatment administration

In the next sections, we will start with CSL to learn a significant causal diagram from the data and then quantify the effect of treatment (IV-Input) on the outcome (mortality status, denoted by the Died_within_48H variable) through CEL. Finally, we use CPL to find the best individualized treatment rules under different settings.

10.1.1 CSL on MIMIC-III. For our analysis of the MIMIC-III dataset, we employ a score-based method in CSL to estimate the underlying causal relationships among several key clinical features. The MIMIC-III dataset comprises a comprehensive range of clinical data from a large cohort of ICU patients. For this analysis, we select a subset of variables, including Glucose levels, PaO2/FiO2 ratio, SOFA scores, IV-Input, and mortality within 48 hours. The details of the selected features are pivotal for understanding patient outcomes in intensive care, and their descriptions are provided in Table 10. Specifically, we utilize the NOTEARS algorithm (X. Zheng, Aragam, et al. 2018), which is designed to learn a DAG from continuous data by enforcing acyclicity through a smooth constraint, given the complexity of the observed data.

The resulting DAG obtained from the NOTEARS algorithm is presented in Figure 7, which reveals a plausible causal structure among the variables. In particular, PaO2/FiO2, a measure of lung efficiency, is identified as an

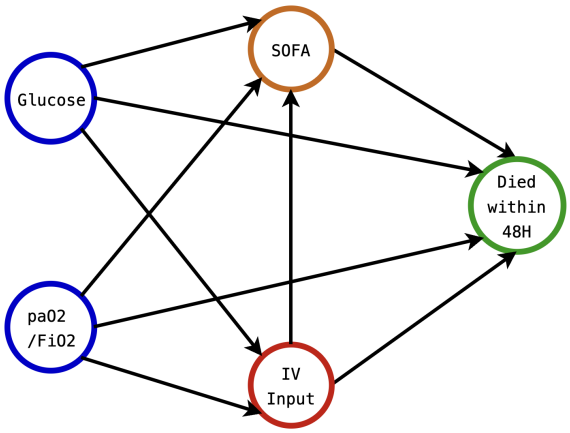


Fig. 7. Estimated directed acyclic graph for the selected MIMIC-III data analysis.

exogenous variable that influences other downstream variables but is not influenced by any of the variables selected in this analysis. Glucose levels and IV-Input appear causally prior to other variables, suggesting their role in early medical interventions. The SOFA score, a critical measure of organ failure, is influenced by both glucose levels and IV-Input and, in turn, impacts the mortality outcome. The mortality within 48 hours variable is positioned as an endpoint in the causal chain, influenced by the SOFA score.

The learned causal graph highlights several clinically relevant pathways. Notably, the direct influence of IV-Input on SOFA score and mortality illustrates the impact of fluid management in critical care settings. Additionally, the model suggests a pathway from metabolic control (Glucose) through organ function (SOFA score) to mortality outcomes. These findings could inform targeted interventions aimed at optimizing patient care and improving survival outcomes in the ICU.

10.1.2 CEL on MIMIC-III. After establishing the causal diagram among the relevant variables, the next step is to quantify the causal effect of IV-Input on the mortality rate of patients. For simplicity, we categorize the treatment, specifically, IV-Input, into a binary action space: $A_i = 1$ represents the "High-IV-Input" group with IV-Input volume greater than 1, and $A_i = 0$ corresponds to the "Low-IV-Input" group with IV-Input less than or equal to 1. Our goal is to apply CEL methods to estimate the ATE, quantifying the impact as $\mathbb{E}\{R(1) - R(0)\}$.

In the CSL analysis, we examined the causal relationships among the variables listed in Table 10 and identified a mediator, the SOFA score, which is influenced by the treatment (IV-Input) and subsequently impacts the mortality status of patients within the next 48 hours. Utilizing the direct estimator proposed by Robins and Greenland (1992), the IPW estimator from Hong et al. (2010), and the robust estimator introduced by Tchetgen and Shpitser (2012), we evaluated the natural direct and indirect effects of the treatment regimen based on observational data. The results are summarized in Table 11 below.

Table 11. Comparison of DE, IE, and TE across different estimation methods.

Methods	**DE**	**IE**	**TE**
Direct Estimator	-0.2133	0.0030	-0.2104
Inverse Probability Weighting	-0.2332	0	-0.2332
Doubly Robust	-0.2276	-0.0164	-0.2440

Specifically, compared to the lower IV-Input group, constantly administering a high volume of IV-Input showed a negative impact on survival rates, with an estimated $20\% - 25\%$ increase in mortality. This effect is primarily driven by the direct influence of the treatment on the final outcome. While this result may seem counterintuitive at first, it could be partly attributed to the potential overuse of IV-Input in clinical practice.

Therefore, developing a personalized treatment regimen that optimizes IV-Input volume according to individual patient characteristics is crucial to avoid overuse and meet specific treatment needs. This challenge motivates our exploration of offline policy learning in the next section.

10.1.3 Offline CPL on MIMIC-III. In this section, we demonstrate the results of applying classic offline CPL methods to this dataset. The simplest modeling approach usually starts from paradigm 1, where we aggregate the observations for each patient and use the average observations as the dataset for the analysis.

As an example, we use the Q-learning algorithm to evaluate two simple treatment rules based on the observed data. We specify the following linear model as our Q-function:

$$\begin{aligned} Q(s, a, \beta) = \beta_{00} + \beta_{01} \cdot \text{Glucose} + \beta_{02} \cdot \text{PaO2\_FiO2} \\ + I(a_1 = 1) \cdot (\beta_{10} + \beta_{11} \cdot \text{Glucose} + \beta_{12} \cdot \text{PaO2\_FiO2}) \end{aligned}$$

We evaluate two target policies. The first is a fixed treatment regimen that does not apply treatment, for which Q-learning has an estimated value of .99. Another is a fixed treatment regimen that applies treatment all the time, with an estimated value of .76. Therefore, the result implies that a high dose of IV-Input naively is always worse and increases the mortality rate, which aligns with the CEL results.

We take one step further to find an optimal policy maximizing the expected value. We use the Q-learning algorithm again to perform policy optimization. Using the regression model specified above, the estimated optimal policy is to recommend $A = 0$ (IV-Input = 0) if $-0.0003 \cdot \text{Glucose} + 0.0012 \cdot \text{PaO2\_FiO2} < 0.5633$ and $A = 1$ (IV-Input = 1) otherwise. When applying the estimated optimal treatment regimen to individuals in the observed data, IV-Input would be administered to 6 out of the 57 patients.

Based on domain knowledge, it is usually more plausible to believe the outcome of a patient depends only on the patient's treatment and condition in past stages. Therefore, we can apply a 3-stage Q-learning in Paradigm 3 to learn the policy. The Q-function is specified as linear, considering all previous stages' states and actions. The learned optimal policy is as follows:

- **Stage 1:** recommend $A = 0$ (IV-Input = 0) if $0.0001 \cdot \text{Glucose}_1 + 0.0012 \cdot \text{PaO2\_FiO2}_1 > 0.0551$ and $A = 1$(IV-Input = 1) otherwise.
- **Stage 2:** recommend $A = 0$ (IV-Input = 0) if $0.0002 \cdot \text{Glucose}_2 - 0.00001 \cdot \text{PaO2\_FiO2}_2 + 0.0070 \cdot \text{SOFA}_1 < 0.0721$ and $A = 1$ (IV-Input = 1) otherwise.
- **Stage 3:** recommend $A = 0$ (IV-Input = 0) if $-0.0005 \cdot \text{Glucose}_2 + 0.0008 \cdot \text{PaO2\_FiO2}_2 - 0.0114 \cdot \text{SOFA}_2 < 0.2068$ and recommend $A = 1$ (IV-Input = 1) otherwise.

Applying the estimated optimal regimen to individuals in the observed data yields personalized treatment plans. For example, 23 patients will receive IV-Input in the first two stages and no inputs in the last one, while 10 others will receive IV-Input only in the first stage.

10.2 MovieLens

Recommender systems play a crucial role in personalizing user experiences across various industries. A common example is movie recommendation, where understanding user preferences across different movie genres is essential. However, this task is challenging due to the inherent difficulty of estimating counterfactuals, i.e., how users would have responded if presented with different options. To illustrate how recommender systems can be optimized through causal decision-making, we use movie recommendations as an example, starting with the well-known MovieLens dataset.

The MovieLens 1M dataset, derived from an online movie recommendation experiment, is a widely used benchmark for online bandit simulation studies. User information in this dataset is categorized by age, gender, and occupation. For simplicity, we focus on the top five movie genres in the dataset (Comedy $a = 0$, Drama $a = 1$, Action $a = 2$, Thriller $a = 3$, Sci-Fi $a = 4$) and analyze users from the top five occupations. The realized reward R takes values in $\{1, 2, 3, 4, 5\}$, with 1 indicating the lowest and 5 the highest satisfaction level. As the causal structure of this problem has been well defined, with movie genre as the action and the user's rating as the reward, our objectives are twofold, mainly focusing on CEL and CPL.

First, we begin the process with CEL, where scientists analyze the logged data to identify general patterns of user preferences. Specifically, the ATE of one movie genre relative to another is calculated to reveal overarching trends across the user population, and HTEs are estimated to capture variations in preferences across different user segments, providing a more granular understanding of how different groups respond to various types of content. These insights lay the groundwork for CPL.

Second, given the dynamic nature of user preferences and frequent interactions between users and the system, movie recommendation is typically approached as an online CPL problem. The primary challenge is balancing the exploitation of existing knowledge about user preferences with the need to explore new data to improve counterfactual estimations. We will detail in this section that offline counterfactual estimates obtained through CEL offer valuable guidance for managing the exploration-exploitation trade-off in the early stages of online policy learning by informing data collection strategies. By further employing diverse online CPL methodologies, the recommender system can dynamically adapt and refine its recommendations, leading to a more personalized and optimal user experience.

To simulate a real-world recommender system, we randomly sample 1% of the dataset to serve as the logged data currently available for offline analysis, while using the entire dataset to estimate the true reward distribution, which will be used to simulate the observed rewards during online interactions.

10.2.1 CEL on MovieLens. In CEL, we aim to estimate the potential outcomes (i.e., the expected ratings) of users across different movie genres. Using the T-learner approach, we fit a separate regression model for each genre (arm) to estimate the expected rating for each individual user. Table 12 below provides a summary of the expected ratings for two subgroups, Female and Male, across these different movie genres.

Table 12. Expected ratings of movie genres for different gender group

Genre	Expected Rating (Female)	Expected Rating (Male)
Comedy	3.580	3.445
Drama	3.403	3.424
Action	3.282	3.073
Thriller	3.512	3.236
Sci-Fi	3.082	2.958

In Table 12, we can see that except for Drama, females tend to give higher expected ratings than males across the various movie genres. Depending on the researcher's objectives, CEL approaches, including both ATE and HTE estimators provided in Section 6.2, offer multiple avenues to understand individual preferences across movie genres.

10.2.2 Online CPL on MovieLens. Movie recommendation has been extensively studied as an online bandit problem (paradigm 5), with its continuous feasibility of data collection. In this section, we simulate a real-world movie recommendation system using the full MovieLens dataset to demonstrate the necessity of online learning in decision-making and to further illustrate how insights from CEL can be applied to CPL. As an example, we

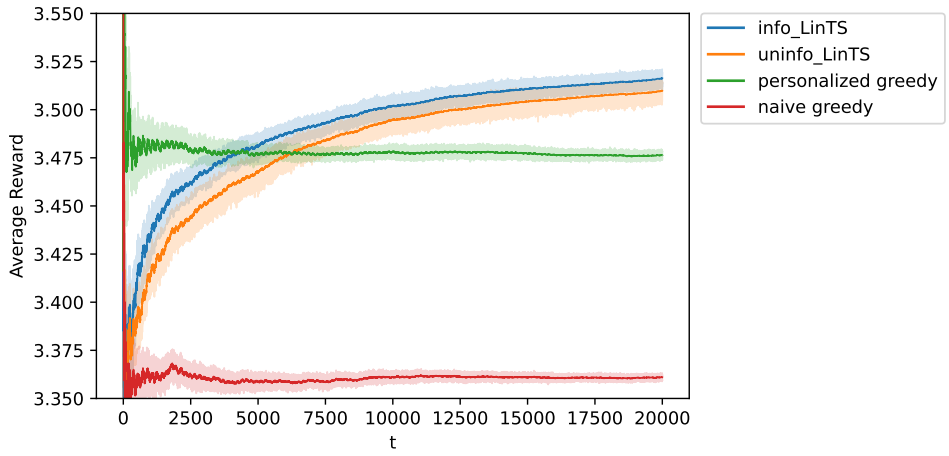


Fig. 8. Simulation results for movie recommendation. Shaded areas indicate the 95% confidence interval of the averages over 50 replicates.

implement the Linear Thompson Sampling (LinTS) algorithm to learn the optimal policy online. Specifically, we assume that for each arm, $R_t(a) \sim \mathcal{N}(s_a^T\gamma, \sigma^2)$, where s_a is a vector contains feature information for the movie genre a, γ is a vector of parameters, and σ^2 is the variance of the random noise. Using 1% of the logged data and the estimates from the CEL step, we first estimated σ and γ. These estimates were then used to construct an informative LinTS, with $\mathcal{N}(\hat{\gamma}, 0.05\boldsymbol{I})$ as the prior distribution for γ, where $\boldsymbol{I}$ is the identity matrix and 0.05 is the prior variance selected to reflect a reasonable confidence in the estimated $\hat{\gamma}$ from the logged data while acknowledging the remaining uncertainty that requires further exploration.

To highlight the advantages of incorporating information from the CEL step, we also implement an uninformative LinTS using $\mathcal{N}(\mathbf{0}, 1000\boldsymbol{I})$ as the prior for γ. Additionally, in CEL, we observed that both female and male users, on average, prefer thriller movies with the highest expected ratings. Based on this insight, a simple greedy algorithm recommending thrillers to all users is considered as a baseline. We also implement a personalized greedy algorithm that generates tailored recommendations derived from more granular, individual-level estimates produced by CEL. The simulation results are presented in Figure 8.

Overall, as expected, the naive greedy algorithm, which ignores personalized preferences, performs the worst. While the personalized greedy algorithm outperforms the LinTS algorithms in the early stages due to less random exploration, both LinTS algorithms continue to gather new information from the environment and eventually surpass the performance of the personalized greedy algorithm. Furthermore, when comparing the uninformative LinTS to the informative LinTS, it is evident that the latter performs better—especially in the early stages—thanks to the prior knowledge acquired from the logged data and the CEL step.

10.3 Software Packages for CDM

Before concluding the real data analysis, we provide a summary of existing software packages for causal decision making in Table 13. For each package, we list its name, the specific tasks and paradigms it addresses, and the corresponding reference links. This overview serves as a structured guide for practitioners, facilitating the identification and utilization of relevant tools within the proposed CDM framework.

Table 13. Software Packages for Causal Decision Making

Package	Task	Focus	Link
causal-learn	CSL	Constraint-based and score-based causal discovery methods (e.g., PC, GES)	https://github.com/cmu-phil/causal-learn
GCastle	CSL	Unified framework for structure learning (PC, GES, NOTEARS, DAG-GNN, etc.)	https://github.com/huawei-noah/trustworthyAI/tree/main/gcastle
cdt (Causal Discovery Toolbox)	CSL	Graphical, functional, and neural network-based causal discovery	https://github.com/FenTechSolutions/CausalDiscoveryToolbox
pcalg (R package)	CSL	R package for constraint-based methods (PC, FCI)	https://cran.r-project.org/web/packages/pcalg/
bnlearn(R package)	CSL	Bayesian network structure learning using score- and constraint-based methods	https://www.bnlearn.com/
CausalML	CEL & CPL	OPE in Paradigm 1	https://causalml.readthedocs.io/
EconML	CEL	HTE Estimation for Paradigm 1	https://github.com/py-why/EconML
Dowhy	CEL	ATE, HTE, and Mediation analysis for Paradigm 1	https://github.com/py-why/dowhy
D3RLpy	CEL & CPL	Paradigm 2	https://github.com/takuseno/d3rlpy
CORL (Clean Offline RL)	CEL & CPL	Paradigm 2	https://corl-team.github.io/CORL/
RLlib	CEL & CPL	Paradigm 2 & 5 & 6	https://docs.ray.io/en/latest/rllib/index.html
CausalPy	CEL	Paradigm 1 & 3	https://github.com/pymc-labs/CausalPy
OBP (Open Bandit Pipeline)	CPL	OPO/OPE for Paradigm 1; Paradigm 4	https://github.com/st-tech/zr-obp
SCOPE-RL	CPL	OPO/OPE	https://scope-rl.readthedocs.io/en/latest/
offline.rl.ope	CPL	OPO/OPE for Paradigm 1	https://github.com/st-tech/zr-obp
Caltech OPE Benchmarking Suite (COBS)	CPL	OPE	https://github.com/clvoloshin/COBS
RLKit	CPL	OPO for Paradigm 2	https://github.com/vitchyr/rlkit
DynTxRegime	CPL	OPO/OPE for Paradigm 3 (DTR)	https://cran.r-project.org/package=DynTxRegime
SMPyBandits	CPL	Paradigm 4	https://github.com/SMPyBandits/SMPyBandits

Table 13 – continued

Package	Task	Focus	Link
MABWiser	CPL	Paradigm 4	https://github.com/fidelity/mabwiser?tab=readme-ov-file
ACME	CPL	Paradigm 2 & 5	https://github.com/google-deepmind/acme/tree/master
TnsorForce	CPL	Paradigm 5	https://github.com/tensorforce/tensorforce?tab=readme-ov-file
POMDP-PY	CPL	Paradigm 6	https://github.com/h2r/pomdp-py?tab=readme-ov-file

11 Discussion: Other Open Directions

Beyond the major ongoing research areas in CDM discussed above, several additional directions have also attracted growing attention. These efforts move beyond pure reward maximization as the sole objective. For example, recent studies have increasingly investigated how CSL and CEL techniques can enhance decision making by incorporating additional objectives such as **fairness** and **explainability**.

Motivated by raising awareness of potential discrimination issues, which is essential in building a trustworthy recommendation system, SCM is widely used to help understand the **fairness issues**. J. Zhang and Bareinboim (2018a) decomposes the effect of natural variations in features and adopts the SCM to infer and distinguish different types of natural discrimination; W. Huang et al. (2022) evaluates the counterfactual effect of sensitive attributes on rewards and limits the action space to arms satisfying counterfactual fairness constraints; and Balakrishnan et al. (2022) defines a Path-specific Counterfactual Effect (PCE) to quantify the causal effect of a protected attribute on the reward through a specific path and formulates the fairness-aware recommendation problem as a constrained MDP problem.

Causal knowledge is also useful in enhancing the **explanability** of decision making. Madumal et al. (2020) introduced an action-influence model that captures the causal relationships among variables using structural equations. By continuously learning the SCM, they provide insights into the behavior of RL agents by generating explanations for "why A" and "why not A" questions through counterfactual reasoning based on the learned SCM. Instead of focusing on explaining a single action choice, Tsirtsis et al. (2021) aims to explain observed sequences of multiple interdependent actions. In scenarios involving multiple agents or more complex environments, using SCM for counterfactual reasoning, Triantafyllou et al. (2022) investigated multi-agent RL to disentangle the contributions of individual agents, while Mesnard et al. (2021) differentiated the effect of actions from that of external factors on future rewards. These interconnected topics not only highlight their synergy within causal decision making, but also pave the way for exciting future research directions.

12 Conclusion

Causality seeks to explain *how actions lead to effects*, while decision making focuses on *how to take actions that yield the greatest effects*. In this paper, we present a comprehensive framework for decision making through a causal lens. We decompose CDM into three key tasks (CSL, CEL, CPL) and six paradigms (distinguished by differences in data structures and by offline/online learning settings), with each accompanied by a detailed review of state-of-the-art methods. We take an affirmative step toward highlighting the widespread use of causality in decision making by integrating all three tasks into a unified framework (see Section 5-8), with an extra emphasis on assumption-violated scenarios (see Section 9). To provide a hands-on tutorial, we develope a GitHub notebook and a Python package that summarizes popular methods for each task (CSL, CEL, CPL), which are widely used in real-world applications. Combined with the real-data applications discussed in Section 10, we believe that this

paper offers a comprehensive tutorial for practitioners interested in the intersection of causality and decision making.

Acknowledgments

This work does not relate to the positions at Amazon.

References

A. Abadie and J. Gardeazabal. 2003. "The economic costs of conflict: A case study of the Basque Country." *American economic review*, 93, 1, 113–132.

A. Abadie and G. W. Imbens. 2011. "Bias-corrected matching estimators for average treatment effects." *Journal of Business & Economic Statistics*, 29, 1, 1–11.

A. Abadie and G. W. Imbens. 2006. "Large sample properties of matching estimators for average treatment effects." *econometrica*, 74, 1, 235–267.

A. Abadie and G. W. Imbens. 2016. "Matching on the estimated propensity score." *Econometrica*, 84, 2, 781–807.

A. Abadie and G. W. Imbens. 2008. "On the failure of the bootstrap for matching estimators." *Econometrica*, 76, 6, 1537–1557.

A. Agarwal, A. Agarwal, L. Masoero, and J. Whitehouse. 2024. "Mutli-Armed Bandits with Network Interference." *Advances in Neural Information Processing Systems*, 37, 36414–36437.

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. 2014. "Taming the monster: A fast and simple algorithm for contextual bandits." In: *International Conference on Machine Learning*. PMLR, 1638–1646.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. 2017. "Thompson sampling for the mnl-bandit." In: *Conference on learning theory*. PMLR, 76–78.

S. Agrawal and N. Goyal. 2013a. "Further optimal regret bounds for thompson sampling." In: *Artificial intelligence and statistics*. PMLR, 99–107.

S. Agrawal and N. Goyal. 2013b. "Thompson sampling for contextual bandits with linear payoffs." In: *International conference on machine learning*. PMLR, 127–135.

D. F. Alwin and R. M. Hauser. 1975. "The decomposition of effects in path analysis." *American sociological review*, 37–47.

W. An and T. J. VanderWeele. 2022. "Opening the blackbox of treatment interference: Tracing treatment diffusion through network analysis." *Sociological Methods & Research*, 51, 1, 141–164.

J. D. Angrist and G. W. Imbens. 1995. "Two-stage least squares estimation of average causal effects in models with variable treatment intensity." *Journal of the American statistical Association*, 90, 430, 431–442.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association*, 91, 434, 444–455.

J. D. Angrist and J.-S. Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

J. Antonelli, M. Cefalu, N. Palmer, and D. Agniel. 2018. "Doubly robust matching estimators for high dimensional confounding adjustment." *Biometrics*, 74, 4, 1171–1179.

D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. 2021. "Synthetic difference-in-differences." *American Economic Review*, 111, 12, 4088–4118.

D. Arkhangelsky and G. Imbens. 2024. "Causal models for longitudinal and panel data: a survey." *The Econometrics Journal*, 27, 3, 1–61.

P. M. Aronow and C. Samii. 2017. "Estimating average causal effects under general interference, with application to a social network experiment." *The Annals of Applied Statistics*, 11, 4, 1912.

K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. 2017. "A brief survey of deep reinforcement learning." *arXiv preprint arXiv:1708.05866*. (Cited by 1161 times, as of May 2025).

C. K. Assaad, E. Devijver, and E. Gaussier. 2022. "Survey and evaluation of causal discovery methods for time series." *Journal of Artificial Intelligence Research*, 73, 767–819.

S. Athey. 2019. "21. The Impact of Machine Learning on Economics." In: *The economics of artificial intelligence*. University of Chicago Press, 507–552.

S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. 2021. "Matrix completion methods for causal panel data models." *Journal of the American Statistical Association*, 116, 536, 1716–1730.

S. Athey, R. Chetty, and G. Imbens. 2020. "Combining experimental and observational data to estimate treatment effects on long term outcomes." *arXiv preprint arXiv:2006.09676*. (Cited by 119, as of May 2025).

S. Athey and G. W. Imbens. 2015. "Machine learning methods for estimating heterogeneous causal effects." *stat*, 1050, 5, 1–26.

S. Athey, G. W. Imbens, and S. Wager. 2018. "Approximate residual balancing: debiased inference of average treatment effects in high dimensions." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80, 4, 597–623.

S. ATHEY, J. TIBSHIRANI, and S. WAGER. 2019. "GENERALIZED RANDOM FORESTS." *The Annals of Statistics*, 47, 2, 1148–1178.

P. Auer, N. Cesa-Bianchi, and P. Fischer. 2002. "Finite-time analysis of the multiarmed bandit problem." *Machine learning*, 47, 2, 235–256.
P. C. Austin. 2008. "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." *Statistics in medicine*, 27, 12, 2037–2049.
Y. Bai, Y. Gao, R. Wan, S. Zhang, and R. Song. 2024. "A Review of Reinforcement Learning in Financial Applications." *Annual Review of Statistics and Its Application*, 12.
L. C. Baird. 1994. "Reinforcement learning in continuous time: Advantage updating." In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. Vol. 4. IEEE, 2448–2453.
S. Balakrishnan, J. Bi, and H. Soh. 2022. "SCALES: From Fairness Principles to Constrained Decision-Making." In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 46–55.
H. Bang and J. M. Robins. 2005. "Doubly robust estimation in missing data and causal inference models." *Biometrics*, 61, 4, 962–973.
E. Bareinboim, A. Forney, and J. Pearl. 2015. "Bandits with unobserved confounders: A causal approach." *Advances in Neural Information Processing Systems*, 28.
F. J. Bargagli-Stoffi, C. Tortù, and L. Forastiere. 2025. "Heterogeneous treatment and spillover effects under clustered network interference." *The Annals of Applied Statistics*, 19, 1, 28–55.
E. Bargiacchi, T. Verstraeten, D. Roijers, A. Nowé, and H. Hasselt. 2018. "Learning to coordinate with coordination graphs in repeated single-stage multi-agent decision problems." In: *International conference on machine learning*. PMLR, 482–490.
R. M. Baron and D. A. Kenny. 1986. "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." *Journal of personality and social psychology*, 51, 6, 1173.
M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. 2016. "Unifying count-based exploration and intrinsic motivation." *Advances in neural information processing systems*, 29.
D. Benkeser, M. Carone, M. V. D. Laan, and P. B. Gilbert. 2017. "Doubly robust nonparametric inference on the average treatment effect." *Biometrika*, 104, 4, 863–880.
A. Bibaut, M. Dimakopoulou, N. Kallus, A. Chambaz, and M. van der Laan. 2021. "Post-contextual-bandit inference." *Advances in Neural Information Processing Systems*, 34.
A. Bietti, A. Agarwal, and J. Langford. 2021. "A contextual bandit bake-off." *Journal of Machine Learning Research*, 22, 133, 1–49.
C. R. Blyth. 1972. "On Simpson's paradox and the sure-thing principle." *Journal of the American Statistical Association*, 67, 338, 364–366.
K. A. Bollen. 1987. "Total, direct, and indirect effects in structural equation models." *Sociological methodology*, 37–69.
D. Bouneffouf, I. Rish, and C. Aggarwal. 2020. "Survey on applications of multi-armed and contextual bandits." In: *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 1–8.
J. E. Brand, X. Zhou, and Y. Xie. 2023. "Recent Developments in Causal Inference and Machine Learning." *Annual Review of Sociology*, 49.
P. Bühlmann, J. Peters, J. Ernest, et al.. 2014. "CAM: Causal additive models, high-dimensional order search and penalized regression." *The Annals of Statistics*, 42, 6, 2526–2556.
H. Cai, R. Song, and W. Lu. 2020. "ANOCE: Analysis of Causal Effects with Multiple Mediators via Constrained Structural Learning." In: *International Conference on Learning Representations*.
M. Caliendo and S. Kopeinig. 2008. "Some practical guidance for the implementation of propensity score matching." *Journal of economic surveys*, 22, 1, 31–72.
L. Canese, G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, M. Re, and S. Spanò. 2021. "Multi-agent reinforcement learning: A review of challenges and applications." *Applied Sciences*, 11, 11, 4948.
A. G. Carranza, S. K. Krishnamurthy, and S. Athey. 2023. "Flexible and efficient contextual bandits with heterogeneous treatment effect oracles." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 7190–7212.
A. Chakrabortty, P. Nandy, and H. Li. 2018. "Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models." *arXiv preprint arXiv:1809.10652*. (Cited by 24, as of May 2025).
B. Chakraborty and E. E. Moodie. 2013. "Statistical methods for dynamic treatment regimes." *Springer-Verlag. doi*, 10, 978-1, 4–1.
B. Chakraborty and S. A. Murphy. 2014. "Dynamic treatment regimes." *Annual review of statistics and its application*, 1, 1, 447–464.
A. Chambaz, W. Zheng, and M. J. van der Laan. 2017. "Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward." *Annals of statistics*, 45, 6, 2537.
O. Chapelle and L. Li. 2011. "An empirical evaluation of thompson sampling." *Advances in neural information processing systems*, 24, 2249–2257.
G. Chen, D. Zeng, and M. R. Kosorok. 2016. "Personalized dose finding using outcome weighted learning." *Journal of the American Statistical Association*, 111, 516, 1509–1521.
H. Chen, W. Lu, and R. Song. 2020. "Statistical inference for online decision making: In a contextual bandit setting." *Journal of the American Statistical Association*, 1–16.
L. Chen, C. Li, X. Shen, and W. Pan. 2024. "Discovery and inference of a causal network with hidden confounding." *Journal of the American Statistical Association*, 119, 548, 2572–2584.
M. Chen, Y. Li, E. Wang, Z. Yang, Z. Wang, and T. Zhao. 2021. "Pessimism meets invariance: Provably efficient offline mean-field multi-agent RL." *Advances in Neural Information Processing Systems*, 34, 17913–17926.

S. Chen and B. Zhang. 2023. "Estimating and improving dynamic treatment regimes with a time-varying instrumental variable." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85, 2, 427–453.
W. Chen, Y. Wang, and Y. Yuan. 2013. "Combinatorial multi-armed bandit: General framework and applications." In: *International conference on machine learning*. PMLR, 151–159.
V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *Econom. J.*, 21, 1, C1–C68. doi:10.1111/ectj.12097.
V. Chernozhukov, M. Demirer, G. Lewis, and V. Syrgkanis. 2019. "Semi-parametric efficient policy learning with continuous actions." *Advances in Neural Information Processing Systems*, 32.
D. M. Chickering. 2002. "Optimal structure identification with greedy search." *Journal of machine learning research*, 3, Nov, 507–554.
W. Chu, L. Li, L. Reyzin, and R. Schapire. 2011. "Contextual bandits with linear payoff functions." In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 208–214.
K. Chua, R. Calandra, R. McAllister, and S. Levine. 2018. "Deep reinforcement learning in a handful of trials using probabilistic dynamics models." *Advances in neural information processing systems*, 31.
R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. 2009. "Dealing with limited overlap in estimation of average treatment effects." *Biometrika*, 96, 1, 187–199.
Y. Cui and E. Tchetgen Tchetgen. 2021. "A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity." *Journal of the American Statistical Association*, 116, 533, 162–173.
A. Curth and M. van der Schaar. 2021. "Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 1810–1818.
A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. 2021. "Overlap in observational studies with high-dimensional covariates." *Journal of Econometrics*, 221, 2, 644–654.
B. Dai, O. Nachum, Y. Chow, L. Li, C. Szepesvári, and D. Schuurmans. 2020. "Coindice: Off-policy confidence interval estimation." *Advances in neural information processing systems*, 33, 9398–9411.
R. H. Dehejia and S. Wahba. 1999. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs." *Journal of the American statistical Association*, 94, 448, 1053–1062.
M. Deisenroth and C. E. Rasmussen. 2011. "PILCO: A model-based and data-efficient approach to policy search." In: *Proceedings of the 28th International Conference on machine learning (ICML-11)*, 465–472.
Z. Deng, J. Jiang, G. Long, and C. Zhang. 2023. "Causal Reinforcement Learning: A Survey." *Transactions on Machine Learning Research*. Survey Certification. https://openreview.net/forum?id=qqnttX9LPo.
Y. Deshpande, L. Mackey, V. Syrgkanis, and M. Taddy. 2018. "Accurate inference for adaptive linear models." In: *International Conference on Machine Learning*. PMLR, 1194–1203.
M. Dimakopoulou, Z. Ren, and Z. Zhou. 2021. "Online multi-armed bandits with adaptive inference." *Advances in Neural Information Processing Systems*, 34, 1939–1951.
M. Dimakopoulou, Z. Zhou, S. Athey, and G. Imbens. 2019. "Balanced linear contextual bandits." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, 3445–3453.
P. Ding and F. Li. 2018. "Causal inference." *Statistical Science*, 33, 2, 214–237.
J. Dorresteijn, F. Visseren, P. Ridker, A. Wassink, N. Paynter, E. Steyerberg, Y. van der Graaf, and N. Cook. 2011. "Estimating treatment effects for individual patients based on the results of randomised clinical trials." *BMJ: British medical journal/British Medical Association*, 343, 7828.
A. Dubey et al.. 2020. "Kernel methods for cooperative multi-agent contextual bandits." In: *International Conference on Machine Learning*. PMLR, 2740–2750.
M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. 2011. "Efficient optimal learning for contextual bandits." In: *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011*, 169.
M. Dudík, J. Langford, and L. Li. 2011. "Doubly robust policy evaluation and learning." In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 1097–1104.
A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau. 2018. "Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis." In: *Machine learning for healthcare conference*. PMLR, 67–82.
D. Ernst, P. Geurts, and L. Wehenkel. 2005. "Tree-based batch mode reinforcement learning." *Journal of Machine Learning Research*, 6.
A. Farahat and M. C. Bailey. 2012. "How effective is targeted advertising?" In: *Proceedings of the 21st international conference on World Wide Web*, 111–120.
A. Feder et al.. 2022. "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond." *Transactions of the Association for Computational Linguistics*, 10, 1138–1158.
L. Forastiere, E. M. Airoldi, and F. Mealli. 2021. "Identification and estimation of treatment and interference effects in observational studies on networks." *Journal of the American Statistical Association*, 116, 534, 901–918.
S. Fujimoto, D. Meger, and D. Precup. 2019. "Off-policy deep reinforcement learning without exploration." In: *International conference on machine learning*. PMLR, 2052–2062.

R. Gallop, D. S. Small, J. Y. Lin, M. R. Elliott, M. Joffe, and T. R. Ten Have. 2009. "Mediation analysis with principal stratification." *Statistics in medicine*, 28, 7, 1108–1130.
C. Gao, Y. Zheng, W. Wang, F. Feng, X. He, and Y. Li. 2024. "Causal inference in recommender systems: A survey and future directions." *ACM Transactions on Information Systems*, 42, 4, 1–32.
L. Ge, J. Wang, C. Shi, Z. Wu, and R. Song. 2023. "A reinforcement learning framework for dynamic mediation analysis." In: *International Conference on Machine Learning*. PMLR, 11050–11097.
C. W. J. Granger. 1969. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica*, 37, 3, 424–438.
S. Gronauer and K. Diepold. 2022. "Multi-agent deep reinforcement learning: a survey." *Artificial Intelligence Review*, 55, 2, 895–943.
S. Gruber and M. J. van der Laan. 2010. "A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome." *The International Journal of Biostatistics*, 6, 1.
S. Gruber and M. Van Der Laan. 2012. "tmle: an R package for targeted maximum likelihood estimation." *Journal of Statistical Software*, 51, 1–35.
S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll. 2024. "A review of safe reinforcement learning: Methods, theories and applications." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
S. Gu, T. Lillicrap, Z. Ghahramani, R. E. Turner, and S. Levine. 2017. "Q-Prop: Sample-Efficient Policy Gradient with An Off-Policy Critic." In: *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net.
V. Hadad, D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey. 2021. "Confidence intervals for policy evaluation in adaptive experiments." *Proceedings of the national academy of sciences*, 118, 15, e2014602118.
B. B. Hansen. 2008. "The prognostic analogue of the propensity score." *Biometrika*, 95, 2, 481–488.
K. Harris, A. Agarwal, C. Podimata, and Z. S. Wu. 2024. "Strategyproof decision-making in panel data settings and beyond." *ACM SIGMETRICS Performance Evaluation Review*, 52, 1, 69–70.
N. Harris and M. Drton. 2013. "PC algorithm for nonparanormal graphical models." *Journal of Machine Learning Research*, 14, 11.
U. Hasan, E. Hossain, and M. O. Gani. 2023. "A Survey on Causal Discovery Methods for I.I.D. and Time Series Data." *Transactions on Machine Learning Research*. Survey Certification. https://openreview.net/forum?id=YdMrdhGx9y.
N. Hassanpour and R. Greiner. 2019. "Learning disentangled representations for counterfactual regression." In: *International Conference on Learning Representations*.
M. Hausknecht and P. Stone. 2015. "Deep recurrent q-learning for partially observable mdps." In: *2015 aaai fall symposium series*.
A. F. Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.
J. J. Heckman, H. Ichimura, and P. Todd. 1998. "Matching as an econometric evaluation estimator." *The review of economic studies*, 65, 2, 261–294.
M. A. Hernán and J. M. Robins. 2006. "Estimating causal effects from epidemiological data." *Journal of Epidemiology & Community Health*, 60, 7, 578–586.
M. A. Hernán. 2004. "A definition of causal effect for epidemiological research." *Journal of Epidemiology & Community Health*, 58, 4, 265–271.
M. Á. Hernán, B. Brumback, and J. M. Robins. 2000. "Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men." *Epidemiology*, 561–570.
K. Hirano, G. W. Imbens, and G. Ridder. 2003. "Efficient estimation of average treatment effects using the estimated propensity score." *Econometrica*, 71, 4, 1161–1189.
G. Hong et al.. 2010. "Ratio of mediator probability weighting for estimating natural direct and indirect effects." In: *Proceedings of the American Statistical Association, biometrics section*. Alexandria, VA, USA, 2401–2415.
C. Hsiao. 2022. "Analysis of panel data." *Econometric Society monographs*.
C. Hsiao. 2007. "Panel data analysis—advantages and challenges." *Test*, 16, 1, 1–22.
X. Hu et al.. 2022. "Causality-driven hierarchical structure discovery for reinforcement learning." *Advances in Neural Information Processing Systems*, 35, 20064–20076.
Y. Hu and N. Kallus. 2020. "Dtr bandit: Learning to make response-adaptive decisions with low regret." *arXiv preprint arXiv:2005.02791*. (Cited by 5, as of May 2025).
B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. 2018. "Generalized score functions for causal discovery." In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1551–1560.
W. Huang, L. Zhang, and X. Wu. 2022. "Achieving counterfactual fairness for causal bandit." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36, 6952–6959.
M. G. Hudgens and M. E. Halloran. 2008. "Toward causal inference with interference." *Journal of the American Statistical Association*, 103, 482, 832–842.
K. Imai, L. Keele, and D. Tingley. 2010. "A general approach to causal mediation analysis." *Psychological methods*, 15, 4, 309.
K. Imai, L. Keele, and T. Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science*, 25, 1, 51–71.

G. Imbens, N. Kallus, X. Mao, and Y. Wang. 2024. "Long-term causal inference under persistent confounding via data combination." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkae095.
G. W. Imbens. 2004. "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and statistics*, 86, 1, 4–29.
G. W. Imbens and D. B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.
G. W. Imbens and J. M. Wooldridge. 2009. "Recent developments in the econometrics of program evaluation." *Journal of economic literature*, 47, 1, 5–86.
N. Jaques, A. Ghandeharioun, J. H. Shen, C. Ferguson, A. Lapedriza, N. Jones, S. Gu, and R. Picard. 2019. "Way off-policy batch deep reinforcement learning of implicit human preferences in dialog." *arXiv preprint arXiv:1907.00456*. (Cited by 387, as of May 2025).
O. Jeunen and B. Goethals. 2021. "Pessimistic reward models for off-policy learning in recommendation." In: *Proceedings of the 15th ACM Conference on Recommender Systems*, 63–74.
S. Jia, P. Frazier, and N. Kallus. 2024. "Multi-Armed Bandits with Interference." *arXiv preprint arXiv:2402.01845*. (Cited by 6, as of May 2025).
C. Jiang, M. P. Wallace, and M. E. Thompson. 2023. "Dynamic treatment regimes with interference." *Canadian Journal of Statistics*, 51, 2, 469–502.
J. Jiang and Z. Lu. 2023. "Offline Decentralized Multi-Agent Reinforcement Learning." In: *ECAI*, 1148–1155.
N. Jiang and L. Li. 2016. "Doubly robust off-policy value evaluation for reinforcement learning." In: *International Conference on Machine Learning*. PMLR, 652–661.
Y. Jin, Z. Ren, Z. Yang, and Z. Wang. 2023. "Policy learning "without" overlap: Pessimism and generalized empirical Bernstein's inequality." In: *2023 IMS International Conference on Statistics and Data Science (ICSDS)*, 367.
F. Johansson, U. Shalit, and D. Sontag. 2016. "Learning representations for counterfactual inference." In: *International conference on machine learning*. PMLR, 3020–3029.
A. E. Johnson et al.. 2016. "MIMIC-III, a freely accessible critical care database." *Scientific data*, 3, 1, 1–9.
J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva. 2022. "Causal machine learning: A survey and open problems." *arXiv preprint arXiv:2206.15475*. (Cited by 227, as of May 2025).
M. Kalisch and P. Bühlmann. 2007. "Estimating high-dimensional directed acyclic graphs with the PC-algorithm." *Journal of Machine Learning Research*, 8, Mar, 613–636.
N. Kallus. 2018. "Instrument-armed bandits." In: *Algorithmic Learning Theory*. PMLR, 529–546.
N. Kallus and M. Uehara. 2022. "Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning." *Operations Research*.
N. Kallus and A. Zhou. 2018. "Policy evaluation and optimization with continuous treatments." In: *International conference on artificial intelligence and statistics*. PMLR, 1243–1251.
S. Kanase et al.. 2022. "An Application of Causal Bandit to Content Optimization." In: *Proceedings of the 5th Workshop on Online Recommender Systems and User Modeling co-located with the 16th ACM Conference on Recommender Systems, ORSUM@RecSys 2022, Seattle, WA, USA, September 23rd, 2022* (CEUR Workshop Proceedings). Ed. by J. Vinagre, M. Al-Ghossein, A. M. Jorge, A. Bifet, and L. Peska. Vol. 3303. CEUR-WS.org. https://ceur-ws.org/Vol-3303/paper3.pdf.
K. Keith, D. Jensen, and B. O'Connor. 2020. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
E. H. Kennedy et al.. 2020. "Optimal doubly robust estimation of heterogeneous causal effects." *arXiv preprint arXiv:2004.14497*, 5.
E. H. Kennedy. 2023. "Towards optimal doubly robust estimation of heterogeneous causal effects." *Electronic Journal of Statistics*, 17, 2, 3008–3049.
K. Khamaru, Y. Deshpande, L. Mackey, and M. J. Wainwright. 2021. "Near-optimal inference in adaptive linear regression." *arXiv preprint arXiv:2107.02266*. (Cited by 28, as of May 2025).
S. Khan, M. Saveski, and J. Ugander. 2024. "Off-policy evaluation beyond overlap: Sharp partial identification under smoothness." In: *Forty-first International Conference on Machine Learning*.
G.-S. Kim and M. C. Paik. 2019. "Doubly-robust lasso bandit." *Advances in Neural Information Processing Systems*, 32.
W. Kim, G.-S. Kim, and M. C. Paik. 2021. "Doubly robust thompson sampling with linear payoffs." *Advances in Neural Information Processing Systems*, 34, 15830–15840.
W. Kim, K. Lee, and M. C. Paik. 2023. "Double doubly robust thompson sampling for generalized linear contextual bandits." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37, 8300–8307.
B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez. 2021. "Deep reinforcement learning for autonomous driving: A survey." *IEEE Transactions on Intelligent Transportation Systems*, 23, 6, 4909–4926.
T. Kitagawa and A. Tetenov. 2018. "Who should be treated? empirical welfare maximization methods for treatment choice." *Econometrica*, 86, 2, 591–616.
M. R. Kosorok and E. B. Laber. 2019. "Precision medicine." *Annual review of statistics and its application*, 6, 1, 263–286.

A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine. 2019. "Stabilizing off-policy q-learning via bootstrapping error reduction." *Advances in Neural Information Processing Systems*, 32.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. 2019. "Metalearners for estimating heterogeneous treatment effects using machine learning." *Proceedings of the national academy of sciences*, 116, 10, 4156–4165.

B. Kveton, M. Konobeev, M. Zaheer, C.-w. Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari. 2021. "Meta-thompson sampling." In: *International Conference on Machine Learning*. PMLR, 5884–5893.

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. 2015. "Cascading bandits: Learning to rank in the cascade model." In: *International conference on machine learning*. PMLR, 767–776.

B. Kveton, M. Zaheer, C. Szepesvari, L. Li, M. Ghavamzadeh, and C. Boutilier. 2020. "Randomized exploration in generalized linear bandits." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 2066–2076.

S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien. 2020. "Gradient-Based Neural DAG Learning." In: *International Conference on Learning Representations*.

P. Ladosz, L. Weng, M. Kim, and H. Oh. 2022. "Exploration in deep reinforcement learning: A survey." *Information Fusion*, 85, 1–22.

J. Langford and T. Zhang. 2007. "The epoch-greedy algorithm for contextual multi-armed bandits." *Advances in neural information processing systems*, 20, 1, 96–1.

F. Lattimore, T. Lattimore, and M. D. Reid. 2016. "Causal bandits: Learning good interventions via causal inference." *Advances in neural information processing systems*, 29.

T. Lattimore and C. Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

H. Le, C. Voloshin, and Y. Yue. 2019. "Batch policy learning under constraints." In: *International Conference on Machine Learning*. PMLR, 3703–3712.

F. P. Leacy and E. A. Stuart. 2014. "On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study." *Statistics in medicine*, 33, 20, 3488–3508.

M. Lechner et al.. 2011. "The estimation of causal effects by difference-in-difference methods." *Foundations and Trends® in Econometrics*, 4, 3, 165–224.

S. Lee and E. Bareinboim. 2019. "Structural causal bandits with non-manipulable variables." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, 4164–4172.

S. Lee and E. Bareinboim. 2018. "Structural causal bandits: Where to intervene?" *Advances in neural information processing systems*, 31.

S. Lee and V. Honavar. 2020. "Towards robust relational causal discovery." In: *Uncertainty in Artificial Intelligence*. PMLR, 345–355.

M. P. Leung. 2022a. "Causal inference under approximate neighborhood interference." *Econometrica*, 90, 1, 267–293.

M. P. Leung. 2022b. "Rate-optimal cluster-randomized designs for spatial interference." *The Annals of Statistics*, 50, 5, 3064–3087.

S. Levine, C. Finn, T. Darrell, and P. Abbeel. 2016. "End-to-end training of deep visuomotor policies." *Journal of Machine Learning Research*, 17, 39, 1–40.

S. Levine, A. Kumar, G. Tucker, and J. Fu. 2020. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems." *arXiv preprint arXiv:2005.01643*. (Cited by 2428, as of May 2025).

C. Li, X. Shen, and W. Pan. 2023. "Inference for a large directed acyclic graph with unspecified interventions." *Journal of Machine Learning Research*, 24, 73, 1–48.

C. Li, X. Shen, and W. Pan. 2019. "Likelihood ratio tests for a large directed acyclic graph." *Journal of the American Statistical Association*.

K. T. Li. 2020. "Statistical inference for average treatment effects estimated by synthetic control methods." *Journal of the American Statistical Association*, 115, 532, 2068–2083.

L. Li, W. Chu, J. Langford, and R. E. Schapire. 2010. "A contextual-bandit approach to personalized news article recommendation." In: *Proceedings of the 19th international conference on World wide web*, 661–670.

L. Li, W. Chu, J. Langford, and X. Wang. 2011. "Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms." In: *Proceedings of the fourth ACM international conference on Web search and data mining*, 297–306.

Y. Li, H. Xie, Y. Lin, and J. C. Lui. 2021. "Unifying offline causal inference and online bandit learning for data driven decision." In: *Proceedings of the Web Conference 2021*, 2291–2303.

S. Liang, W. Lu, and R. Song. 2018. "Deep advantage learning for optimal dynamic treatment regime." *Statistical theory and related fields*, 2, 1, 80–88.

L. Liao, Z. Fu, Z. Yang, Y. Wang, D. Ma, M. Kolar, and Z. Wang. 2024. "Instrumental variable value iteration for causal offline reinforcement learning." *Journal of Machine Learning Research*, 25, 303, 1–56.

P. Liao, Z. Qi, R. Wan, P. Klasnja, and S. A. Murphy. 2022. "Batch policy learning in average reward markov decision processes." *Annals of statistics*, 50, 6, 3364.

S.-H. Lin and T. VanderWeele. 2017. "Interventional approach for path-specific effects." *Journal of Causal Inference*, 5, 1.

M. L. Littman. 2009. "A tutorial on partially observable Markov decision processes." *Journal of Mathematical Psychology*, 53, 3, 119–125.

Q. Liu, L. Li, Z. Tang, and D. Zhou. 2018. "Breaking the curse of horizon: Infinite-horizon off-policy estimation." *Advances in Neural Information Processing Systems*, 31.

Y. Liu, Y. Wang, M. R. Kosorok, Y. Zhao, and D. Zeng. 2018. "Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens." *Statistics in medicine*, 37, 26, 3776–3788.

D. Lopez-Paz and B. Schölkopf. 2017. "Discovering Causal Signals in Images." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6979–6987.

Y. Lu, A. Meisami, A. Tewari, and W. Yan. 2020. "Regret analysis of bandit problems with causal background knowledge." In: *Conference on Uncertainty in Artificial Intelligence*. PMLR, 141–150.

S. Luo, Y. Yang, C. Shi, F. Yao, J. Ye, and H. Zhu. 2024. "Policy evaluation for temporal and/or spatial dependent experiments." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkad136.

M. H. Maathuis, M. Kalisch, P. Bühlmann, et al.. 2009. "Estimating high-dimensional intervention effects from observational data." *The Annals of Statistics*, 37, 6A, 3133–3164.

D. P. MacKinnon and J. H. Dwyer. 1993. "Estimating mediated effects in prevention studies." *Evaluation review*, 17, 2, 144–158.

D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz. 2007. "Mediation analysis." *Annu. Rev. Psychol.*, 58, 593–614.

D. P. MacKinnon, C. M. Lockwood, C. H. Brown, W. Wang, and J. M. Hoffman. 2007. "The intermediate endpoint effect in logistic and probit regression." *Clinical trials*, 4, 5, 499–513.

D. P. MacKinnon, C. M. Lockwood, J. M. Hoffman, S. G. West, and V. Sheets. 2002. "A comparison of methods to test mediation and other intervening variable effects." *Psychological methods*, 7, 1, 83.

P. Madumal, T. Miller, L. Sonenberg, and F. Vetere. 2020. "Explainable reinforcement learning through a causal lens." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34, 2493–2500.

C. F. Manski. 2013. "Identification of treatment response with social interactions." *The Econometrics Journal*, 16, 1, S1–S23.

F. Mealli, B. Pacini, and D. B. Rubin. 2011. "Statistical inference for causal effects." *Modern analysis of customer surveys: With applications using R*, 171–192.

J. Mei, Z. Zhong, B. Dai, A. Agarwal, C. Szepesvari, and D. Schuurmans. 2023. "Stochastic gradient succeeds for bandits." In: *International Conference on Machine Learning*. PMLR, 24325–24360.

L. Meng, R. Gorbet, and D. Kulić. 2021. "Memory-based deep reinforcement learning for pomdps." In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5619–5626.

T. Mesnard et al.. 2021. "Counterfactual Credit Assignment in Model-Free Reinforcement Learning." In: *International Conference on Machine Learning*. PMLR, 7654–7664.

A. Miller and K. Hosanagar. 2020. "Personalized discount targeting with causal machine learning." In: *ICIS*.

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. 2016. "Asynchronous methods for deep reinforcement learning." In: *International conference on machine learning*. PMLR, 1928–1937.

V. Mnih, K. Kavukcuoglu, et al.. 2015. "Human-level control through deep reinforcement learning." *Nature*, 518, 7540, 529–533.

T. M. Moerland, J. Broekens, A. Plaat, C. M. Jonker, et al.. 2023. "Model-based reinforcement learning: A survey." *Foundations and Trends® in Machine Learning*, 16, 1, 1–118.

S. L. Morgan and C. Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

S. A. Murphy. 2003. "Optimal dynamic treatment regimes." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65, 2, 331–355.

A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine. 2018. "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning." In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7559–7566.

V. Nair, V. Patil, and G. Sinha. 2021. "Budgeted and non-budgeted causal bandits." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 2017–2025.

P. Nandy, A. Hauser, and M. H. Maathuis. 2018. "High-dimensional consistency in score-based and hybrid structure learning." *The Annals of Statistics*, 46, 6A, 3151–3183.

P. Nandy, M. H. Maathuis, T. S. Richardson, et al.. 2017. "Estimating the effect of joint interventions from observational data in sparse high-dimensional settings." *The Annals of Statistics*, 45, 2, 647–674.

M. Nauta, D. Bucur, and C. Seifert. 2019. "Causal Discovery with Attention-based Convolutional Neural Networks." *Machine Learning and Knowledge Extraction*, 1, 1, 312–340.

B. Neal. 2020. "Introduction to causal inference from a machine learning perspective." *Course Lecture Notes (draft)*. (Cited by 8, as of May 2025).

S. Neel and A. Roth. 2018. "Mitigating bias in adaptive data gathering via differential privacy." In: *International Conference on Machine Learning*. PMLR, 3720–3729.

X. Nie, C. Lu, and S. Wager. 2021. "Nonparametric heterogeneous treatment effect estimation in repeated cross sectional designs." In: *Handbook of Statistical Methods for Precision Medicine*. Chapman and Hall/CRC, 181–206.

X. Nie, X. Tian, J. Taylor, and J. Zou. 2018. "Why adaptively collected data have negative bias and how to correct for it." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 1261–1269.

X. Nie and S. Wager. 2021. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika*, 108, 2, 299–319.

Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen. 2021. "Counterfactual vqa: A cause-effect look at language bias." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12700–12710.
I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. 2016. "Deep exploration via bootstrapped DQN." *Advances in neural information processing systems*, 29.
C. Paduraru. 2013. "Off-policy evaluation in Markov decision processes." (Cited by 44, as of May 2025).
L. Pan, L. Huang, T. Ma, and H. Xu. 2022. "Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification." In: *International conference on machine learning*. PMLR, 17221–17237.
U. Panizza and A. F. Presbitero. 2014. "Public debt and economic growth: is there a causal effect?" *Journal of Macroeconomics*, 41, 21–41.
J. Pearl. 1995. "Causal diagrams for empirical research." *Biometrika*, 82, 4, 669–688.
J. Pearl. 2010a. "Causal inference." *Causality: objectives and assessment*, 39–58.
J. Pearl. 2009. "Causal inference in statistics: An overview." *Statistics surveys*, 3, 96–146.
J. Pearl. 2000. *Causality: models, reasoning and inference*. Vol. 29. Springer.
J. Pearl. 2022. "Direct and indirect effects." In: *Probabilistic and causal inference: the works of Judea Pearl*. ACM, 373–392.
J. Pearl. 2003. "Statistics and causal inference: A review." *Test*, 12, 281–345.
J. Pearl. 2010b. "The foundations of causal inference." *Sociological Methodology*, 40, 1, 75–149.
J. Peters and P. Bühlmann. 2014. "Identifiability of Gaussian structural equation models with equal error variances." *Biometrika*, 101, 1, 219–228.
J. Peters, D. Janzing, and B. Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.
J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. 2014. "Causal discovery with continuous additive noise models." *The Journal of Machine Learning Research*, 15, 1, 2009–2053.
M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. Van Der Laan. 2012. "Diagnosing and responding to violations in the positivity assumption." *Statistical methods in medical research*, 21, 1, 31–54.
R. F. Prudencio, M. R. Maximo, and E. L. Colombini. 2023. "A survey on offline reinforcement learning: Taxonomy, review, and open problems." *IEEE Transactions on Neural Networks and Learning Systems*.
H. Qiu, M. Carone, and A. Luedtke. 2022. "Individualized treatment rules under stochastic treatment cost constraints." *Journal of causal inference*, 10, 1, 480–493.
H. Qiu, M. Carone, E. Sadikova, M. Petukhova, R. C. Kessler, and A. Luedtke. 2021. "Optimal individualized decision rules using instrumental variable methods." *Journal of the American Statistical Association*, 116, 533, 174–191.
P. Ramprasad, Y. Li, Z. Yang, Z. Wang, W. W. Sun, and G. Cheng. 2023. "Online bootstrap inference for policy evaluation in reinforcement learning." *Journal of the American Statistical Association*, 118, 544, 2901–2914.
J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. 2017. "A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images." *International journal of data science and analytics*, 3, 2, 121–129.
P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. 2021. "Bridging offline reinforcement learning and imitation learning: A tale of pessimism." *Advances in Neural Information Processing Systems*, 34, 11702–11716.
J. M. Robins. 2004. "Optimal structural nested models for optimal sequential decisions." In: *Proceedings of the Second Seattle Symposium in Biostatistics: analysis of correlated data*. Springer, 189–326.
J. M. Robins and S. Greenland. 1992. "Identifiability and exchangeability for direct and indirect effects." *Epidemiology*, 3, 2, 143–155.
J. M. Robins, D. Lin, and P. Heagerty. 2004. "Proceedings of the second Seattle symposium in biostatistics." *Optimal Structural Nested Models for Optimal Sequential Decisions*, 189–326.
P. M. Robinson. 1988. "Root-N-consistent semiparametric regression." *Econometrica: Journal of the Econometric Society*, 931–954.
P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello. 2022. "Score matching enables causal discovery of nonlinear additive noise models." In: *International Conference on Machine Learning*. PMLR, 18741–18753.
P. R. Rosenbaum and D. B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70, 1, 41–55.
D. L. Roth and D. P. MacKinnon. 2013. "Mediation analysis with longitudinal data." In: *Longitudinal data analysis*. Routledge, 181–216.
D. B. Rubin. 1978. "Bayesian inference for causal effects: The role of randomization." *The Annals of statistics*, 6, 34–58. ISBN: 0090-5364.
D. B. Rubin. 2004. "Direct and indirect causal effects via potential outcomes." *Scandinavian Journal of Statistics*, 31, 2, 161–170.
D. B. Rubin. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, 66, 5, 688.
G. A. Rummery and M. Niranjan. 1994. *On-line Q-learning using connectionist systems*. Vol. 37. University of Cambridge, Department of Engineering Cambridge, UK.
J. Runge. 2018. "Causal network reconstruction from time series: From theoretical assumptions to practical estimation." *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28, 7.
J. Runge et al.. 2019. "Detecting and quantifying causal associations in large nonlinear time series datasets." *Science Advances*, 5, 11, eaau4996.

A. Sauter, N. Botteghi, E. Acar, and A. Plaat. 2024. "CORE: Towards Scalable and Efficient Causal Discovery with Reinforcement Learning." In: *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 1664–1672.

F. Sävje, P. Aronow, and M. Hudgens. 2021. "Average treatment effects in the presence of unknown interference." *Annals of statistics*, 49, 2, 673.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. 2021. "Toward causal representation learning." *Proceedings of the IEEE*, 109, 5, 612–634.

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. 2015. "Trust region policy optimization." In: *International conference on machine learning*. PMLR, 1889–1897.

J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. 2015. "High-dimensional continuous control using generalized advantage estimation." *arXiv preprint arXiv:1506.02438*. (Cited by 4668, as of May 2025).

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. "Proximal policy optimization algorithms." *arXiv preprint arXiv:1707.06347*. (Cited by 26760, as of May 2025).

P. J. Schulte, A. A. Tsiatis, E. B. Laber, and M. Davidian. 2014. "Q-and A-learning methods for estimating optimal dynamic treatment regimes." *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29, 4, 640.

M. Seitzer, B. Schölkopf, and G. Martius. 2021. "Causal influence detection for improving efficiency in reinforcement learning." *Advances in Neural Information Processing Systems*, 34, 22905–22918.

J. P. Selig and K. J. Preacher. 2009. "Mediation models for longitudinal data in developmental research." *Research in human development*, 6, 2-3, 144–164.

R. Sen, K. Shanmugam, A. G. Dimakis, and S. Shakkottai. 2017. "Identifying best interventions through online importance sampling." In: *International Conference on Machine Learning*. PMLR, 3057–3066.

R. Sen, K. Shanmugam, M. Kocaoglu, A. Dimakis, and S. Shakkottai. 2017. "Contextual bandits with latent confounders: An nmf approach." In: *Artificial Intelligence and Statistics*. PMLR, 518–527.

A. K. Shakya, G. Pillai, and S. Chakrabarty. 2023. "Reinforcement learning algorithms: A brief survey." *Expert Systems with Applications*, 231, 120495.

U. Shalit, F. D. Johansson, and D. Sontag. 2017. "Estimating individual treatment effect: generalization bounds and algorithms." In: *International conference on machine learning*. PMLR, 3076–3085.

W. Shen, J. Wang, Y.-G. Jiang, and H. Zha. 2015. "Portfolio choices with orthogonal bandit learning." In: *Twenty-fourth international joint conference on artificial intelligence*.

Y. Shen, H. Cai, and R. Song. 2024. "Doubly robust interval estimation for optimal policy evaluation in online learning." *Journal of the American Statistical Association*, 119, 548, 2811–2821.

C. Shi, A. Fan, R. Song, and W. Lu. 2018. "High-dimensional A-learning for optimal dynamic treatment regimes." *Annals of statistics*, 46, 3, 925.

C. Shi and L. Li. 2021. "Testing mediation effects using logic of Boolean matrices." *Journal of the American Statistical Association*, 1–14.

C. Shi, S. Luo, Y. Le, H. Zhu, and R. Song. 2024. "Statistically efficient advantage learning for offline reinforcement learning in infinite horizons." *Journal of the American Statistical Association*, 119, 545, 232–245.

C. Shi, M. Uehara, J. Huang, and N. Jiang. 2022. "A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes." In: *International Conference on Machine Learning*. PMLR, 20057–20094.

C. Shi, R. Wan, V. Chernozhukov, and R. Song. 2021. "Deeply-debiased off-policy interval estimation." In: *International Conference on Machine Learning*. PMLR, 9580–9591.

C. Shi, R. Wan, G. Song, S. Luo, H. Zhu, and R. Song. 2023. "A multiagent reinforcement learning framework for off-policy evaluation in two-sided markets." *The Annals of Applied Statistics*, 17, 4, 2701–2722.

C. Shi, X. Wang, S. Luo, H. Zhu, J. Ye, and R. Song. 2023. "Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework." *Journal of the American Statistical Association*, 118, 543, 2059–2071.

C. Shi, D. Blei, and V. Veitch. 2019. "Adapting neural networks for the estimation of treatment effects." *Advances in neural information processing systems*, 32.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. 2006. "A linear non-Gaussian acyclic model for causal discovery." *Journal of Machine Learning Research*, 7, Oct, 2003–2030.

J. Shin, A. Ramdas, and A. Rinaldo. 2019. "Are sample means in multi-armed bandits positively or negatively biased?" *Advances in Neural Information Processing Systems*, 32.

J. Shin, A. Ramdas, and A. Rinaldo. 2021. "On the bias, risk, and consistency of sample means in multi-armed bandits." *SIAM Journal on Mathematics of Data Science*, 3, 4, 1278–1300.

N. Siegel, J. T. Springenberg, F. Berkenkamp, A. Abdolmaleki, M. Neunert, T. Lampe, R. Hafner, N. Heess, and M. Riedmiller. 2020. "Keep Doing What Worked: Behavior Modelling Priors for Offline Reinforcement Learning." In: *International Conference on Learning Representations*.

N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha. 2022. "Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions." *Expert Systems with Applications*, 197, 116669.

D. Silver, T. Hubert, et al.. 2018. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science*, 362, 6419, 1140–1144.

D. Silver, J. Schrittwieser, et al.. 2017. "Mastering the game of Go with deep neural networks and tree search." *Nature*, 550, 354–359.
D. Simchi-Levi and C. Wang. 2023. "Multi-armed bandit experimental design: Online decision-making and adaptive inference." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 3086–3097.
B. Singh, R. Kumar, and V. P. Singh. 2022. "Reinforcement learning in robotic applications: a comprehensive survey." *Artificial Intelligence Review*, 55, 2, 945–990.
S. P. Singh and R. S. Sutton. 1996. "Reinforcement learning with replacing eligibility traces." *Machine learning*, 22, 1-3, 123–158.
A. Slivkins et al.. 2019. "Introduction to multi-armed bandits." *Foundations and Trends® in Machine Learning*, 12, 1-2, 1–286.
M. E. Sobel. 1987. "Direct and indirect effects in linear structural equation models." *Sociological Methods & Research*, 16, 1, 155–176.
M. E. Sobel. 2006. "What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference." *Journal of the American Statistical Association*, 101, 476, 1398–1407.
R. Song, M. Kosorok, D. Zeng, Y. Zhao, E. Laber, and M. Yuan. 2015. "On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning." *Stat*, 4, 1, 59–68.
R. Song, W. Wang, D. Zeng, and M. R. Kosorok. 2015. "Penalized q-learning for dynamic treatment regimens." *Statistica Sinica*, 25, 3, 901.
M. T. Spaan. 2012. "Partially observable Markov decision processes." *Reinforcement learning: State-of-the-art*, 387–414.
P. Spirtes, C. Glymour, R. Scheines, S. Kauffman, V. Aimale, and F. Wimberly. 2000. "Constructing Bayesian network models of gene expression networks from microarray data." (Cited by 206, as of May 2025).
P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. 2000. *Causation, prediction, and search*. MIT press.
E. A. Stuart. 2010. "Matching methods for causal inference: A review and a look forward." *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25, 1, 1.
L. Su, W. Lu, and R. Song. 2019. "Modelling and estimation for optimal treatment decision with interference." *Stat*, 8, 1, e219.
Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudík. 2020. "Doubly robust off-policy evaluation with shrinkage." In: *International Conference on Machine Learning*. PMLR, 9167–9176.
C. Subramanian and B. Ravindran. 2021. "Causal contextual bandits with targeted interventions." In: *International Conference on Learning Representations*.
R. S. Sutton. 1991. "Dyna, an integrated architecture for learning, planning, and reacting." *ACM Sigart Bulletin*, 2, 4, 160–163.
R. S. Sutton. 1988. "Learning to predict by the methods of temporal differences." *Machine learning*, 3, 1, 9–44.
R. S. Sutton and A. G. Barto. 1999. "Reinforcement learning: An introduction." *Robotica*, 17, 2, 229–235.
R. S. Sutton and A. G. Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, and I. Zitouni. 2017. "Off-policy evaluation for slate recommendation." In: *Advances in Neural Information Processing Systems*, 3632–3642.
K. Tang, J. Huang, and H. Zhang. 2020. "Long-tailed classification by keeping the good and removing the bad momentum causal effect." *Advances in neural information processing systems*, 33, 1513–1524.
A. Tank, E. B. Fox, and A. Shojaie. 2021. "Neural Granger Causality for Nonlinear Time Series." *Artificial Intelligence*, 297, 103502.
E. J. T. Tchetgen and I. Shpitser. 2012. "Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis." *Annals of statistics*, 40, 3, 1816.
S. Triantafyllou, A. Singla, and G. Radanovic. 2022. "Actual causality and responsibility attribution in decentralized partially observable Markov decision processes." In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 739–752.
D. Tschernutter, T. Hatt, and S. Feuerriegel. 2022. "Interpretable off-policy learning via hyperbox search." In: *International Conference on Machine Learning*. PMLR, 21795–21827.
A. A. Tsiatis, M. Davidian, S. T. Holloway, and E. B. Laber. 2019. *Dynamic treatment regimes: Statistical methods for precision medicine*. Chapman and Hall/CRC.
S. Tsirtsis, A. De, and M. Rodriguez. 2021. "Counterfactual explanations in sequential decision making under uncertainty." *Advances in Neural Information Processing Systems*, 34, 30127–30139.
M. Uehara, J. Huang, and N. Jiang. 2020. "Minimax weight and q-function learning for off-policy evaluation." In: *International Conference on Machine Learning*. PMLR, 9659–9668.
M. Uehara, C. Shi, and N. Kallus. 2022. "A review of off-policy evaluation in reinforcement learning." *arXiv preprint arXiv:2212.06355*. (Cited by 93, as of May 2025).
T. VanderWeele. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.
T. VanderWeele and S. Vansteelandt. 2014. "Mediation analysis with multiple mediators." *Epidemiologic methods*, 2, 1, 95–115.
T. J. VanderWeele. 2016. "Mediation analysis: a practitioner's guide." *Annual review of public health*, 37, 1, 17–32.
T. J. VanderWeele. 2008. "Simple relations between principal stratification and direct and indirect effects." *Statistics & Probability Letters*, 78, 17, 2957–2962.
T. J. VanderWeele, J. W. Jackson, and S. Li. 2016. "Causal inference and longitudinal data: a case study of religion and mental health." *Social psychiatry and psychiatric epidemiology*, 51, 1457–1466.

T. J. VanderWeele and W. R. Robinson. 2014. "On causal interpretation of race in regressions adjusting for confounding and mediating variables." *Epidemiology (Cambridge, Mass.)*, 25, 4, 473.
T. J. VanderWeele and S. Vansteelandt. 2009. "Conceptual issues concerning mediation, interventions and composition." *Statistics and its Interface*, 2, 4, 457–468.
K. Vermeulen and S. Vansteelandt. 2016. "Data-adaptive bias-reduced doubly robust estimation." *The international journal of biostatistics*, 12, 1, 253–282.
T. Verstraeten, E. Bargiacchi, P. J. Libin, J. Helsen, D. M. Roijers, and A. Nowé. 2020. "Multi-agent thompson sampling for bandit applications with sparse neighbourhood structures." *Scientific reports*, 10, 1, 6728.
G. Visconti and J. R. Zubizarreta. 2018. "Handling limited overlap in observational studies with cardinality matching." *Observational Studies*, 4, 1, 217–249.
D. Viviano. 2024. "Policy targeting under network interference." *Review of Economic Studies*, rdae041.
D. Viviano and J. Bradic. 2022. "Synthetic learner: model-free inference on treatments over time." *Journal of Econometrics*.
D. Viviano and J. Bradic. 2023. "Synthetic learner: model-free inference on treatments over time." *Journal of Econometrics*, 234, 2, 691–713.
C. Voloshin, H. M. Le, and Y. Yue. 2019. "Empirical analysis of off-policy policy evaluation for reinforcement learning." In: *Real-world Sequential Decision Making Workshop at ICML*. Vol. 2019.
M. J. Vowels, N. C. Camgoz, and R. Bowden. 2021. "D'ya like DAGs? A survey on structure learning and causal discovery." *ACM Computing Surveys (CSUR)*.
S. Wager and S. Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*, 113, 523, 1228–1242.
R. Wan, L. Ge, and R. Song. 2021. "Metadata-based multi-task bandits with bayesian hierarchical models." *Advances in Neural Information Processing Systems*, 34, 29655–29668.
R. Wan, L. Ge, and R. Song. 2023. "Towards scalable and robust structured bandits: A meta-learning framework." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 1144–1173.
R. Wan, H. Wei, B. Kveton, and R. Song. 2023. "Multiplier bootstrap-based exploration." In: *International Conference on Machine Learning*. PMLR, 35444–35490.
R. Wan, S. Zhang, C. Shi, S. Luo, and R. Song. 2021. "Pattern transfer learning for reinforcement learning in order dispatching." *RLITS Workshop, IJCAI 2021*.
L. Wang and E. Tchetgen Tchetgen. 2018. "Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80, 3, 531–550.
Y.-X. Wang, A. Agarwal, and M. Dudık. 2017. "Optimal and adaptive off-policy evaluation in contextual bandits." In: *International Conference on Machine Learning*. PMLR, 3589–3597.
X. Wang, S. Wang, X. Liang, D. Zhao, J. Huang, X. Xu, B. Dai, and Q. Miao. 2022. "Deep reinforcement learning: A survey." *IEEE Transactions on Neural Networks and Learning Systems*, 35, 4, 5064–5078.
C. J. Watkins and P. Dayan. 1992. "Q-learning." *Machine learning*, 8, 3-4, 279–292.
R. A. Watson, H. Cai, X. An, S. Mclean, and R. Song. 2023. "On heterogeneous treatment effects in heterogeneous causal graphs." In: *International Conference on Machine Learning*. PMLR, 36714–36747.
R. J. Williams. 1992. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." *Reinforcement learning*, 8, 3-4, 5–32.
P. G. Wright. 1928. *The tariff on animal and vegetable oils*. 26. Macmillan.
L. Wu and S. Yang. 2022. "Integrative *R*-learner of heterogeneous treatment effects combining experimental and observational studies." In: *Conference on Causal Learning and Reasoning*. PMLR, 904–926.
X. Xiang and S. Foo. 2021. "Recent advances in deep reinforcement learning applications for solving partially observable markov decision processes (pomdp) problems: Part 1—fundamentals and applications in games, robotics and natural language processing." *Machine Learning and Knowledge Extraction*, 3, 3, 554–581.
H. Xu and H. Xie. 2023. "A Thompson Sampling Approach to Unifying Causal Inference and Bandit Learning." In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 255–266.
L. Xu, H. Kanagawa, and A. Gretton. 2021. "Deep proxy causal learning and its application to confounded bandit policy evaluation." *Advances in Neural Information Processing Systems*, 34, 26264–26275.
Y. Xu, W. Lu, and R. Song. 2024. "Linear Contextual Bandits with Interference." *arXiv preprint arXiv:2409.15682*. (Cited by 3, as of May 2025).
Y. Xu, J. Zhu, C. Shi, S. Luo, and R. Song. 2023. "An instrumental variable approach to confounded off-policy evaluation." In: *International Conference on Machine Learning*. PMLR, 38848–38880.
S. Yang, J. K. Kim, and R. Song. 2020. "Doubly robust inference when combining probability and non-probability samples with high dimensional data." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82, 2, 445–465.
S. Yang and Y. Zhang. 2023. "Multiply robust matching estimators of average and quantile treatment effects." *Scandinavian Journal of Statistics*, 50, 1, 235–265.

Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang. 2018. "Mean field multi-agent reinforcement learning." In: *International Conference on Machine Learning*. PMLR, 5571–5580.

Y. Yang, X. Ma, C. Li, Z. Zheng, Q. Zhang, G. Huang, J. Yang, and Q. Zhao. 2021. "Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning." *Advances in Neural Information Processing Systems*, 34, 10299–10312.

L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang. 2021. "A survey on causal inference." *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15, 5, 1–46.

M. Yin and Y.-X. Wang. 2020. "Asymptotically efficient off-policy evaluation for tabular reinforcement learning." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 3948–3958.

Y. Yu, J. Chen, T. Gao, and M. Yu. 2019. "DAG-GNN: DAG structure learning with graph neural networks." In: *International conference on machine learning*. PMLR, 7154–7163.

Y. Yuan, X. Shen, W. Pan, and Z. Wang. 2019. "Constrained likelihood for reconstructing a directed acyclic Gaussian graph." *Biometrika*, 106, 1, 109–125.

Y. Zeng, R. Cai, F. Sun, L. Huang, and Z. Hao. 2024. "A survey on causal reinforcement learning." *IEEE Transactions on Neural Networks and Learning Systems*.

R. Zhan, V. Hadad, D. A. Hirshberg, and S. Athey. 2021. "Off-policy evaluation via adaptive weighting with data from contextual bandits." In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2125–2135.

J. Zhang and E. Bareinboim. 2018a. "Fairness in decision-making—the causal explanation formula." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32.

J. Zhang and E. Bareinboim. 2018b. "Non-parametric path analysis in structural causal models." In: *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*.

J. Zhang and E. Bareinboim. 2017. "Transfer learning in multi-armed bandit: a causal approach." In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 1778–1780.

K. Zhang, L. Janson, and S. Murphy. 2020. "Inference for batched bandits." *Advances in neural information processing systems*, 33, 9818–9829.

K. Zhang, L. Janson, and S. Murphy. 2021. "Statistical inference with m-estimators on adaptively collected data." *Advances in Neural Information Processing Systems*, 34, 7460–7471.

Y. Zhang, E. B. Laber, M. Davidian, and A. A. Tsiatis. 2018. "Interpretable dynamic treatment regimes." *Journal of the American Statistical Association*, 113, 524, 1541–1549.

Y. Zhang, S. Yang, W. Ye, D. E. Faries, I. Lipkovich, and Z. Kadziola. 2022. "Practical recommendations on double score matching for estimating causal effects." *Statistics in medicine*, 41, 8, 1421–1445.

Y. Zhao, M. Goodman, S. Kanase, S. Xu, Y. Kimmel, B. Payne, S. Khan, and P. Grao. 2022. "Mitigating Targeting Bias in Content Recommendation with Causal Bandits." In: *Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems co-located with 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, WA, USA, 18th-23rd September 2022* (CEUR Workshop Proceedings). Ed. by H. Abdollahpouri, S. Sahebi, M. Elahi, M. Mansoury, B. Loni, Z. Nazari, and M. Dimakopoulou. Vol. 3268. CEUR-WS.org. https://ceur-ws.org/Vol-3268/paper11.pdf.

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. 2012. "Estimating individualized treatment rules using outcome weighted learning." *Journal of the American Statistical Association*, 107, 499, 1106–1118.

W. Zheng and M. van der Laan. 2017. "Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes." *Journal of causal inference*, 5, 2, 20160006.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. 2018. "DAGs with no tears: Continuous optimization for structure learning." In: *Advances in Neural Information Processing Systems*, 9472–9483.

X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing. 2020. "Learning sparse nonparametric dags." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 3414–3425.

Q. Zhou, X. Zhang, J. Xu, and B. Liang. 2017. "Large-scale bandit approaches for recommender systems." In: *International Conference on Neural Information Processing*. Springer, 811–821.

Y. Zhou, Z. Qi, C. Shi, and L. Li. 2023. "Optimizing Pessimism in Dynamic Treatment Regimes: A Bayesian Learning Approach." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 6704–6721.

J. Zhu, R. Wan, Z. Qi, S. Luo, and C. Shi. 2024. "Robust offline reinforcement learning with heavy-tailed rewards." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 541–549.

S. Zhu, I. Ng, and Z. Chen. 2020. "Causal Discovery with Reinforcement Learning." In: *International Conference on Learning Representations*.

W. Zhu, D. Zeng, and R. Song. 2019. "Proper inference for value function in high-dimensional Q-learning for dynamic treatment regimes." *Journal of the American Statistical Association*, 114, 527, 1404–1417.

A Acronyms

2SLS Two-Stage Least Squares. 39

AIPW augmented inverse probability weighting. 24
ANOCE analysis of causal effects. 19–21
ATE Average Treatment Effect. 7, 10, 11, 22–25, 27, 28, 39, 47
ATT Average Treatment Effect on the Treated. 28

CART classification and regression trees. 32
CDM Causal Decision Making. 2–4, 9, 11, 43, 48, 50
CEL Causal Effect Learning. 3–5, 10, 11, 13, 15, 19, 21–23, 26–31, 34, 35, 39, 40, 44–48, 50
CI conditional independence. 17
CPDAG completed partially directed acyclic graph. 15, 18, 19
CPL Causal Policy Learning. 3–5, 11, 13, 15, 19, 28, 29, 33, 39, 44, 46, 47, 50
CSL Causal Structure Learning. 2–5, 9–20, 26, 35, 44, 45, 50

DAG directed acyclic graph. 7, 15–19, 44
DE direct effect. 8, 19, 41
DiD Difference-in-Difference. 28
DM direct (outcome regression) estimator. 23, 27, 30
DR doubly robust. 23, 25, 27, 30–32, 35
DS double score. 24
DTR Dynamic Treatment Regimes. 3, 5, 11, 26, 32, 34, 40, 41

EHR electronic health record. 43

FQE Fitted-Q Evaluation. 30, 31

HTE Heterogeneous Treatment Effect. 7, 8, 10, 11, 22, 23, 25–28, 39, 47

ICU intensive care unit. 43–45
IE indirect effect. 8, 19
IPW inverse probability weighting. 23, 24, 27, 30–32, 35, 43
IS importance sampling. 23, 31
IV Instrumental Variables. 39, 40

LSEM linear structural equation model. 16–18, 26

MAB Multi-Armed Bandit. 35, 42
MARL Multi-Agent RL. 41, 42
MBRL Model-based reinforcement learning. 32, 33
MDP Markov Decision Process. 3, 5, 6, 11, 12, 26, 27, 29, 31, 34, 39–41, 50
MEC Markov equivalence class. 15, 18, 19
MIMIC-III Medical Information Mart for Intensive Care III. 21, 25, 43–46

NUC No Unmeasured Confounders. 9, 23, 25, 38, 39

OLS ordinary least squares. 15
OPE Off-Policy Evaluation. 5, 28, 29, 31, 32, 40, 42, 43

OPO Off-Policy Optimization. 5, 29, 43

PC Peter-Clark. 17
PCE Path-specific Counterfactual Effect. 50
POMDP Partially Observable Markov Decision Process. 3, 5, 12, 34, 37, 40

RCM Rubin Causal Model. 5
RL Reinforcement Learning. 5, 11, 12, 23, 34, 37–40, 42, 50

SC Synthetic Control. 28
SCM Structural Causal Model. 6, 7, 24, 36, 50
SE spillover effect. 40, 41
SEM Structural Equation Models. 4, 5
SNPs single nucleotide polymorphisms. 14
SOFA Sepsis-related Organ Failure Assessment. 8, 44–46
SUTVA Stable Unit Treatment Value Assumption. 9, 23, 25, 38, 40
SVM support vector machine. 32

TCDF Temporal Causal Discovery Framework. 21
TE total effect. 8, 19
TMLE Targeted Maximum Likelihood Estimation. 24
TS Thompson Sampling. 35–37, 42

UCB Upper Confidence Bound. 35–37, 42

Received 9 November 2025; accepted 10 March 2026