Causal Reinforcement Learning: A Survey

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning is an essential paradigm for solving sequential decision problems under uncertainty. Despite many remarkable achievements in recent decades, applying reinforcement learning methods in the real world remains challenging. One of the main obstacles is that reinforcement learning agents lack common knowledge of the world and must therefore learn from scratch through numerous interactions. They may also struggle to explain their decisions and generalize the learned knowledge. Causality, on the other hand, has a distinct advantage in that it can formalize knowledge and utilize structural invariance for efficient knowledge transfer. This has led to the emergence of causal reinforcement learning, a subfield of reinforcement learning that seeks to improve existing algorithms using structured and interpretable representations of the data generation process. In this survey, we comprehensively review the literature on causal reinforcement learning. We first introduce the basic concepts of causality and reinforcement learning, and then explain how causal modeling can address core challenges in non-causal reinforcement learning. We categorize and systematically review existing causal reinforcement learning approaches based on their target problems and methodologies. Finally, we outline open issues and future directions in this emerging field.

1 Introduction

> *"All reasonings concerning matter of fact seem to be founded on the relation of cause and effect. By means of that relation alone we can go beyond the evidence of our memory and senses."*
>
> *—David Hume, An Enquiry Concerning Human Understanding.*

Humans have an innate ability to develop an understanding of causality from a young age (Wellman, 1992; Inagaki & Hatano, 1993; Koslowski & Masnick, 2002; Sobel & Sommerville, 2010). This level of understanding allows us to realize that changing certain things can cause others to happen; therefore, we can actively intervene in our environment to achieve desired goals or acquire new knowledge. Understanding cause and effect empowers us to explain behavior (Schult & Wellman, 1997), predict the future (Shultz, 1982), and even conduct counterfactual reasoning to reflect on past events (Harris et al., 1996). These abilities are essential to the development of human intelligence, laying the foundation for modern society and civilization as well as advancing science and technology.

For example, consider the story of humans battling scurvy, as illustrated in Figure 1 (Pearl & Mackenzie, 2018). Scurvy once hampered human exploration of the world and claimed the lives of approximately 2 million sailors. After a long quest, humans discovered that consuming citrus fruits could prevent this terrible disease. Today, we know that the actual cause of scurvy is the lack of vitamin C, but at the time, this causal mechanism was unclear. People first believed acidity could cure the disease. However, heating the juice for purification destroyed the vitamin C content, rendering it ineffective against scurvy. People then believed acidity was only a placebo and rotten meat was the cause of the disease. This misguided judgment took a heavy toll on Scott's Antarctica expedition. It was only when the causality of scurvy was fully understood that effective solutions to combat the disease were discovered. This example shows the importance of understanding causality in decision-making and the potentially disastrous consequences of ignoring it.

Figure 1: The journey to discover the cause of scurvy. At first, people thought that acidity could prevent the disease, which led to reckless decision-making, such as heating the juice for purification, which failed to address the root cause of the problem. Later, people started to believe that rotten meat was the cause of scurvy, which is a classic example of mistaking correlation for causality. Finally, it was discovered that the actual cause of scurvy is a deficiency in vitamin C. This breakthrough solved the mystery of scurvy and saved many lives.

Data alone cannot answer causal questions. Understanding causality involves making and testing assumptions about the data generation process. Data-driven machine learning can effectively capture the correlation between citrus fruits and scurvy but cannot handle causality. For example, if we replace citrus fruits with animal livers (also rich in vitamin C) in the scurvy prediction problem, the algorithm will probably give an incorrect prediction due to the significant differences in appearance and taste. Causal machine learning (Schölkopf et al., 2021; Kaddour et al., 2022) was developed to address this deficiency. In recent years, the combination of causality with machine learning has gained much attention and has been applied in various fields, including computer vision (Lopez-Paz et al., 2017; Shen et al., 2018; Tang et al., 2020; Wang et al., 2020b), natural language processing (Wu et al., 2021; Jin et al., 2021; Feder et al., 2022), and recommender systems (Zheng et al., 2021; Zhang et al., 2021b; Gao et al., 2022). These results show that causal modeling significantly improves the distribution robustness and knowledge transferability of learning systems.

Unlike other machine learning paradigms, reinforcement learning (RL) (Sutton & Barto, 2018) involves intervening in the environment to actively collect training data; in this sense, RL is naturally connected to causality. However, in most studies, agents are only allowed to intervene in action variables, making it difficult to fully comprehend causal relationships. This difficulty is further compounded in off-policy and offline settings.

In RL, agents aim to obtain high-return data; thus, they continuously improve their policies by trial and error. In this dynamic process, the environment responds to agents' actions by shifting from the current state to a new one and returning a scalar reward (or penalty). Both the state transition and reward assignment are causal; for example, vitamin C deficiency (the current state) causes scurvy (the next state), but not vice versa. Other environmental factors, such as the flavor and appearance of food, do not affect this transition. To avoid being plagued by non-causal correlations, an agent must capture the causal relationships that drive the underlying data generation process; otherwise, it would learn inefficiently or even get stuck with suboptimal policies.

Many researchers have examined principled ways to integrate causal knowledge with RL. The most popular one is the use of causal graphs (Glymour et al., 2016), a qualitative form of causal knowledge. Causal graphs can be used to represent high-level, coarse-grained data generation processes that do not distinguish the meaning of each dimension, such as a standard Markov decision process (MDP). Meanwhile, causal graphs can also convey low-level, fine-grained causal knowledge, such as decomposing states into multiple variables according to their causal relations. Furthermore, causal knowledge can be represented quantitatively according to the structural causal model (SCM) framework (Pearl, 2009a;b), which we will explain further in section 2. SCM considers the data generation process as an ordered collection of equations that generate data in a structured manner. As we later demonstrate in section 4, RL agents equipped with an SCM can directly generate data without interacting with the actual environment, achieving counterfactual data augmentation and policy evaluation.

Much research has been undertaken in the field of causal RL to explore the various settings, such as bandits and MDPs; online and offline; and model-free and model-based settings. These studies have consistently shown that causal RL algorithms tend to perform better and be more stable than traditional algorithms. However, they are not well connected with each other. Given the central role of causality in human intelligence, causal RL has the potential to overcome many of the limitations of existing methods and to tackle new challenges in more complex decision-making problems. This paper aims to provide a comprehensive survey of causal RL. We establish connections between existing work based on the SCM framework, which allows for the systematic and principal incorporation of causal knowledge into the learning process.

Our main contributions to the field are as follows.

- Our survey of causal RL presents a comprehensive overview of the field, aligning existing research within the SCM framework. In particular, we introduce causal RL by answering three fundamental questions: What it is causal RL? Why does it need to be studied? And how does causal modeling improve existing RL approaches? We also present a clear and concise overview of the foundational concepts of causality research and RL. To the best of our knowledge, this is the first comprehensive survey of causal RL in the extant literature on RL [1].
- We identify the bottleneck problems in RL that can be solved or improved by means of causal modeling. We further propose a problem-oriented taxonomy. This taxonomy will help RL researchers gain a deeper understanding of the advantages of causal modeling and the opportunities for further research. On the other hand, RL practitioners can also benefit from this survey by identifying solutions to the challenges they face. Additionally, we compare and analyze existing causal reinforcement learning research based on their techniques and settings.
- We highlight major unresolved issues and promising research directions in causal RL, such as theoretical advances, benchmarks, and specific learning paradigms. These research topics will become increasingly important in the coming years and will help advance the use of RL in real-world applications. Therefore, having a common ground for discussing these valuable ideas in this emerging field is crucial and will facilitate its continued development and success.

2 Background

To better understand causal RL, an emerging field that combines the strengths of causality research and RL, we start by introducing the fundamentals of and some common concepts relevant to the two research areas.

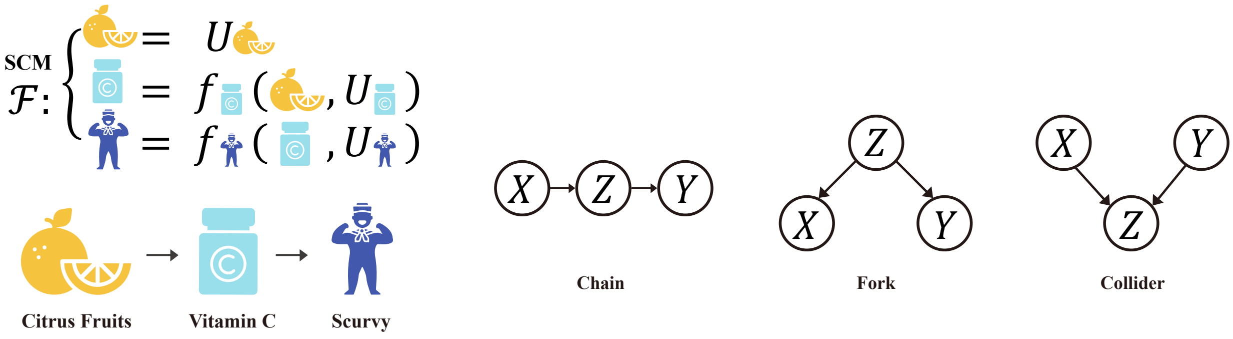
2.1 A Brief Introduction to Causality

We first discuss how to use mathematical language to describe and study causality. In general, there are two primary frameworks that researchers use to formalize causality: SCM (structural causal model) Pearl (2009a); Glymour et al. (2016) and PO (potential outcome) (Rubin, 1974; Imbens & Rubin, 2015). We focus on the former in this paper because it provides a graphical methodology that can help researchers abstract and better understand the data generation process. It is noteworthy that these two frameworks are logically equivalent, and most assumptions are interchangeable.

Definition 2.1 (Structual Causal Model)**.** An SCM is represented by a quadruple $(\mathcal{V}, \mathcal{U}, \mathcal{F}, P(\mathbf{U}))$, where

- $\mathcal{V} = \{V_1, V_2, \cdots, V_n\}$ is a set of endogenous variables that are of interest in a research problem,
- $\mathcal{U} = \{U_1, U_2, \cdots, U_n\}$ is a set of exogenous variables corresponding to $\mathcal{V}$, which are typically unobservable and represent the source of stochasticity, such as noise,
- $\mathcal{F} = \{f_1, f_2, \cdots, f_n\}$ is a set of structural equations that assign values to each of the variables in $\mathcal{V}$,

[1] We note that Schölkopf et al. (2021) and Kaddour et al. (2022) discussed causal RL alongside many other research subjects in their papers. The former mainly studied the causal representation learning problem, and the latter comprehensively investigated the field of causal machine learning. The present study, however, focuses on examining the literature on causal RL and provides a systematic review of the field.



(a) (b)

Figure 2: (a) The SCM and the causal graph of the scurvy prediction problem. In this simplified example, we consider only three endogenous variables and their associated exogenous variables. The nodes corresponding to the exogenous variables are usually omitted in the causal graph. (b) The three basic building blocks of causal graphs.

- $P(\mathbf{U})$ is the joint probability distribution of the exogenous variables in $\mathcal{U}$.

Structural causal model. In definition 2.1, the exogenous variables $\mathbf{U}$ are the input and source of uncertainty for the entire system. In other words, once all exogenous variables are set, all endogenous variables are determined accordingly. Each structural equation $f_i \in \mathcal{F}$ assigns an endogenous variable V_i whose independent variables consist of a corresponding exogenous variable U_i and other endogenous variables (the causes of V_i). In this way, the structural equations mathematically characterize the causality between the variables and the causal mechanisms behind the data generation process. By specifying the distribution of the exogenous variables $P(\mathbf{U})$ and sequentially executing all the structural equations in $\mathcal{F}$, any sample can be generated from the joint distribution $P(\mathbf{V})$.

Causal graph. Each SCM is associated with a causal graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where nodes $\mathcal{V}$ represent endogenous variables and edges $\mathcal{E}$ represent causal relationships determined by the structural equations. In general, causal graphs are directed acyclic graphs. An edge $e_{ij} \in \mathcal{E}$ from node V_j to node V_i exists if the random variable V_j is an independent variable in the structural equation f_{V_i} of variable V_i. In other words, if V_j is part of the structural equation of V_i in the SCM, then it is considered a parent node of V_i on the causal graph, i.e., $V_j \in \mathbf{PA}(V_i)$. Figure 2a shows the SCM of the scurvy problem (a simplified version) and the corresponding causal graph. While structural equations are often unknown in complicated problems, we usually have prior knowledge of causal graphs, which is already sufficient for solving many causal reasoning problems. Figure 2b shows the three basic building blocks of the causal graph, i.e., chain, fork, and collider. These three simple structures can be combined to create more complex data generation processes.

Intervention. Intervention is a way of actively participating in the data generation process, rather than passively observing it. Essentially, interventions create new data distributions, or, in other words, all distributional shifts result from different interventions in the data generation process. There are two types of interventions: One involves directly fixing variables to constant values, which is known as a hard intervention, and the other is retaining some of the original dependencies while making changes to the structural equations (e.g., changing the distribution of exogenous variables), which is called a soft intervention. Many research questions involve predicting the effects of interventions. For example, finding a way to prevent scurvy is essentially about identifying effective interventions (through food or medicine) that lower the probability of getting scurvy. To differentiate from conditional probability, researchers introduced the do-operator, using $P(Y|\mathrm{do}(X) = x)$ to denote the intervention probability, meaning the probability distribution of the outcome variable Y when variable X is fixed to x. Figure 3a illustrates the difference between conditional and intervention probabilities.

Counterfactual. Counterfactual thinking is all about asking the “what if ...?” questions, such as "What if the scurvy patient had eaten enough citrus fruit? Would they stay healthy?". Such a thinking process

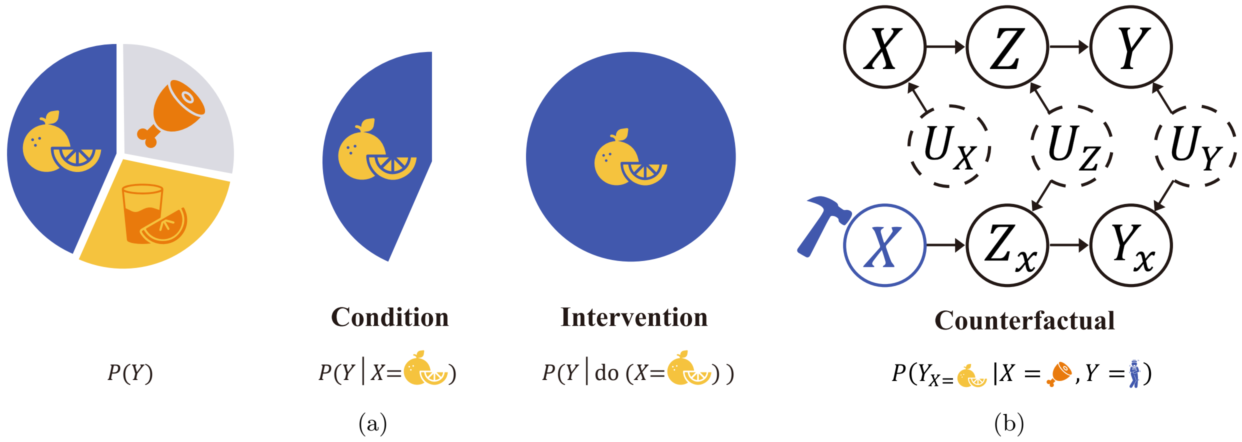


Figure 3: (a) The difference between condition and intervention. Marginal probabilities account for all subgroups within a population, while conditional probabilities focus on a specific subgroup, like sailors who have consumed sufficient citrus fruits. Intervention, on the other hand, looks at the probability of scurvy by forcing all sailors to consume enough citrus fruits per day. (b) An illustration of the counterfactual probability. Counterfactual probability explores events that occur in the imaginary world (the bottom row), such as asking whether sailors would be protected from scurvy if they had consumed enough citrus fruit, given that they consumed meat in the factual world. Note that the exogenous variables in the imaginary world (marked by dashed circles) are consistent with the factual world (the upper row), but intervention (marked by a hammer) may change the outcome of the imaginary world.

emerges in our everyday lives, allowing us to reflect on our behavior in past events and eventually make improvements.

In the research community, counterfactual variables are often denoted with a subscript, such as $Y_{X=x}$ (or Y_x when there is no ambiguity). This approach helps researchers differentiate the counterfactual variables from the original variable Y. Counterfactual reasoning, building on this formalism, aims to estimate probabilities such as $P(Y_{X=1}|X = 0, Y = 1)$. We can consider counterfactual reasoning as creating an imaginary world different from the factual one, whereas intervention only studies the factual one. In particular, if we perform the same intervention in both the factual and imaginary worlds, they overlap - $P(Y_{X=0}|X = 0, Y = 1) = P(Y|\text{do}(X = 0))$. See Figure 3b for a visual representation of counterfactual reasoning.

Causal discovery and causal reasoning. In research on causality, there are two main areas of focus: causal discovery and causal reasoning. Causal discovery involves using data on variables of interest to infer the causal relationships between them (in other words, to identify the causal graph of these variables). Traditional approaches use conditional independence tests to infer causal relationships, and recently some studies have been conducted based on large datasets using deep learning techniques. Glymour et al. (2019) and Vowels et al. (2022) comprehensively surveyed the field of causal discovery.

On the other hand, causal reasoning investigates how to estimate causal effects, such as intervention probability, given the causal model. Interventions involve actively manipulating the system or environment, which can be costly and potentially dangerous (e.g., testing a new drug in medical experiments). Therefore, a core challenge of causal reasoning is how to convert causal effects into statistical estimates that can be inferred from the observed data. Given the causal graph, the identifiability of causal effects can be determined systematically through the use of do-calculus (Pearl, 1995).

Causal factorization. Causal graphs provide a structured factorization approach along causal directions for a joint distribution. This is referred to as causal factorization:

$$P(\mathbf{V}) = \prod_{i=1}^{n} P(V_i|\text{PA}(V_i)), \tag{1}$$

where the (conditional) probability distribution of the form $P(V_i|\mathrm{PA}(V_i))$ is known as the causal mechanism. Causal factorization is unique to the data generation process, but other factorization forms may also satisfy the statistical dependencies manifested by the joint distribution. Consider Figure 2 as an example. The data generation process takes a chain structure. The joint distribution $P(X, Y, Z)$ can be factorized as $P(X)P(Z|X)P(Y|Z)$ or $P(X)P(Y|X)P(Z|X, Y)$. The former is causal, and the latter is not. Causal factorization allows for more efficient learning and inference due to the sparsity of the causal graph structure.

Moreover, causal factorization facilitates generalization to new problems. Consider an intervention concerning vitamin C intake, which alters the joint distribution. An agent with causal factorization only needs to adjust one module ($P(Z|X)$) to predict scurvy under the new distribution, while a non-causal model would have to relearn two modules ($P(Y|X)$ and $P(Z|X, Y)$). This is because the intervention only affects the causal mechanism of vitamin C intake, which is independent of the probability of getting scurvy. This property is referred to as modularity (Pearl, 2009b), or the *independent causal mechanism* (Schölkopf et al., 2021), indicating that the causal generation process comprises a series of stable and autonomous modules (causal mechanisms). Changing one of these modules does not affect the others. Furthermore, small changes in the distribution usually involve only a few modules in causal factorization, providing a principle for designing efficient machine learning algorithms and models.

2.2 A Brief Introduction to Reinforcement Learning

Reinforcement learning studies sequential decision problems. Mathematically, we can formalize these problems as Markov decision processes.

Definition 2.2 (Markov decision process)**.** An MDP $\mathcal{M}$ is specified by a tuple $\{\mathcal{S}, \mathcal{A}, P, R, \mu_0, \gamma\}$, where

- $\mathcal{S}$ denotes the state space and $\mathcal{A}$ denotes the action space,
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function that yields the probability distribution of the next states s_{t+1} after taking an action a_t at the current state s_t,
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function that assigns the immediate reward for taking an action a_t at state s_t,
- $\mu_0 : \mathcal{S} \to [0, 1]$ is the initial probability distribution of states, and
- $\gamma \in [0, 1]$ denotes the discount factor that accounts for how much future events lose their value as time passes.

Markov decision processes. In definition 2.2, the decision process starts by sampling an initial state s_0 with μ_0. An agent takes responsive action using its policy π (a function that maps a state to an action) and receives a reward from the environment assigned by R. The environment evolves to a new state following P; then, the agent senses the new state and repeats the interaction with the environment. The goal of an RL agent is to search for the optimal policy π^* that maximizes the return (cumulative reward) G_0. In particular, at any timestep t, the return G_t is defined as the sum of discounted future rewards, i.e., $G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+i}$. A multi-armed bandit (MAB) is a special type of MDP problem that only considers a single-step decision. A partially observable Markov decision process (POMDP), on the other hand, generalizes the MDP by allowing for partial observability. While the system still operates on the basis of an MDP, the agent must make decisions based on limited information about the system state; for example, in a video game, a player may need to reason about the trajectory of the target object based on what is shown on the screen.

Value functions. Return G_t can evaluate how good an action sequence is. However, when it comes to stochastic environments, the same action sequence can generate different trajectories, resulting in different returns. In particular, a policy π may also be stochastic. Given the current state, a policy π can output a probability distribution over the action space. Thus, return G_t is a random variable. To evaluate a policy, RL introduces the concept of value functions. There are two types of value functions: $V^\pi(s)$ denotes the expected return obtained by following the policy π from state s; $Q^\pi(s, a)$ denotes the expected return obtained by performing action a at state s and following the policy π thereafter. The optimal value functions that correspond to the optimal policy π^* are denoted by $V^*(s)$ and $Q^*(s,a)$.

Bellman equations. By definition, $V^\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$ and $Q^\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$. These two types of value functions can be expressed in terms of one another. By expanding the return G_t, we can rewrite value functions in a recursive manner:

$$\begin{aligned} V^\pi(s) &= \sum_{a\in\mathcal{A}} \pi(a|s) \left(R(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) V^\pi(s') \right) \\ Q^\pi(s,a) &= R(s,a) + \gamma \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} \pi(a'|s') Q^\pi(s',a'). \end{aligned} \tag{2}$$

When the timestep t is not specified, s and s' are often used to refer to the states of two adjacent steps. The above equations are known as the Bellman expectation equations, which establish the connection between two adjacent steps. Similarly, the Bellman optimally equations relate the optimal value functions:

$$\begin{aligned} V^*(s) &= \max_{a\in\mathcal{A}} \left(R(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) V^*(s') \right) \\ Q^*(s,a) &= R(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) \max_{a'\in\mathcal{A}} Q^*(s',a'). \end{aligned} \tag{3}$$

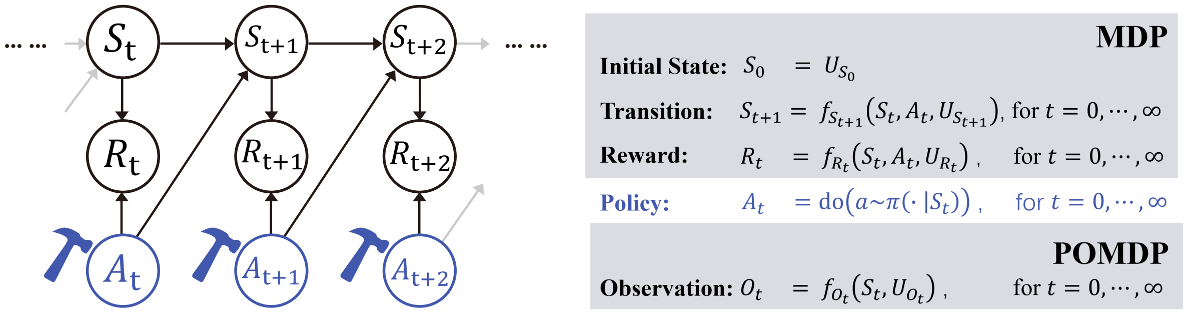
When the environment (also referred to as the dynamic model or, simply, the model) is already known, the learning problem degenerates into a planning problem, which can be solved by dynamic programming based on the Bellman equation. However, RL focuses on the setting of unknown environments, i.e., the agents do not have complete knowledge in $P(s'|s,a)$ and $R(s,a)$, which aligns more closely with decision-making problems under uncertainty in the real world.

Categorizing reinforcement learning methods. There are several ways to categorize RL methods. One way is based on the components of the agent. Policy-based methods focus on optimizing an explicitly parameterized policy to maximize the return, while value-based methods use collected data to fit a value function and derive the policy implicitly from it. Actor-critic methods combine both of them, equipping an agent with a value function and a policy. Another way to classify RL methods is based on whether they use an environmental model. Model-based reinforcement learning (MBRL) typically uses a well-defined environmental model (such as AlphaGo (Silver et al., 2017)) or constructs one using the collected data. The model assists the agent in planning or generating additional training data, thus improving the learning process. Lastly, RL can be divided into on-policy, off-policy, and offline approaches based on how data is collected. On-policy RL only uses data from the current policy, while off-policy RL may involve the data collected by the previous policies. Offline RL disallows data collection, so the agent can only learn from a fixed dataset.

2.3 Causal Reinforcement Learning

An MDP characterizes the data generation process in an RL problem. From a causal perspective, we can reformulate an MDP as an SCM. In particular, we consider the state, action, and reward at each step to be endogenous variables. The state transition and reward functions are then described as deterministic functions with independent exogenous variables, represented by the structural equations $\mathcal{F}$ in the SCM. The initial state can also be considered an exogenous variable such that $P_\mathcal{U}$ includes μ_0. This transformation holds for any MDP and can be achieved by autoregressive uniformization (also known as the reparameterization trick) (Buesing et al., 2019).

The causal graph and the SCM that correspond to an MDP are shown in Figure 4. It is noteworthy that the policy π is not a causal relationship; it is usually a soft intervention that preserves the dependence on the state variables, i.e., $\text{do}(a \sim \pi(\cdot|s))$. As we mentioned previously, interventions create different data distributions; therefore, different policies lead to varying trajectory distributions with different expected returns. It is clear that on-policy RL learns from intervention data, allowing it to estimate the causal effects



MDP

Initial State: $S_0 = U_{S_0}$

Transition: $S_{t+1} = f_{S_{t+1}}(S_t, A_t, U_{S_{t+1}})$, for $t = 0, \cdots, \infty$

Reward: $R_t = f_{R_t}(S_t, A_t, U_{R_t})$, for $t = 0, \cdots, \infty$

Policy: $A_t = \mathrm{do}(a \sim \pi(\cdot \mid S_t))$, for $t = 0, \cdots, \infty$

POMDP

Observation: $O_t = f_{O_t}(S_t, U_{O_t})$, for $t = 0, \cdots, \infty$

Figure 4: A causal perspective on the MDP problem. Action variables are marked with hammers next to them because the policy that generates the actions is an intervention in the data generation process, not a causal relationship. In contrast, the mapping from state to observation in a partially observable Markov decision process (POMDP) problem is a causal relationship.

of actions directly. In contrast, off-policy and offline RL involve the agent passively observing and learning from data collected by previous policies. In this case, the training data is observational from the learner's viewpoint, which may be easily confounded with spurious correlations (Deng et al., 2021).

The SCM framework allows us to discuss causality in decision-making problems, understand how different distributions arise, and organize causal knowledge in a clear and reusable way. Furthermore, SCM allows us to explore counterfactual problems in RL, which is impossible for non-causal approaches. In this paper, we define causal RL as follows.

Definition 2.3 (Causal reinforcement learning)**.** Causal RL is an umbrella term for RL approaches that focus on understanding and utilizing causal mechanisms of the underlying data generation process to inform decision-making.

Definition 2.3 states that causal RL differs from other forms of RL in that it focuses on causality rather than other superficial correlations or patterns in the training data. We remark that in RL, the agent seeks to find the policy with the highest expected return, not just to infer the causal effect of interventions. In the following section, we will demonstrate that causal modeling not only improves the understanding of causal relationships but is also effective in various problems in which non-causal approaches are insufficient.

3 Why Causality Is Important in Reinforcement Learning

Reinforcement learning has made remarkable advancements in the last decade, but it still carries some significant challenges. In this section, we summarize four major obstacles that have hindered the widespread use of RL algorithms in real-world applications, for which causal modeling can offer prospective solutions. We then analyze these challenges and explain why they can benefit from causal modeling.

3.1 Sample Efficiency in Reinforcement Learning

3.1.1 The Issue of Sample Efficiency in Reinforcement Learning

In RL, the data used for training is not provided beforehand. Unlike supervised and unsupervised learning methods that directly learn from a fixed dataset, an RL agent needs to actively gather new data to optimize its policy toward the highest return. An effective RL algorithm should be able to master the optimal policy with as few experiences as possible (in other words, it must be sample-efficient). Current methods often require collecting millions of samples to succeed in even simple tasks, let alone more complicated environments and reward mechanisms. For example, AlphaGo Zero was trained over roughly 3×10^7 games of self-play (Silver et al., 2017); OpenAI's Rubik's Cube robot took nearly 10^4 years of simulation experience (OpenAI et al., 2019). This inefficiency entails a high training costs and prevents the use of RL techniques for solving real-world decision-making problems that change rapidly. Therefore, the sample efficiency issue is a core challenge in RL, and developing sample-efficient RL algorithms that can save time and resources is essential.

3.1.2 Why Causal Modeling Helps Improve Sample Efficiency?

Researchers of RL have long been concerned with sample efficiency (Kakade, 2003; Osband et al., 2013; Grande et al., 2014; Yu, 2018). This issue usually relates to several elements of the learning process. The first one is abstraction, which is also a fundamental problem in machine learning. A properly abstracted problem is easier to solve due to the reduced dimensionality. Understanding the environment and learning policies in the abstraction space can be much more efficient. Some approaches (Jong & Stone, 2005; Zhang et al., 2022) realize abstraction by aggregating multiple states that produce the same reward sequence. Although these methods effectively reduce the dimensionality by eliminating irrelevant variables, there are still redundant dependencies among the remaining variables that are vulnerable to spurious correlations. Causal modeling simplifies the complexity of the problem by identifying causal variables in the environmental dynamics and preserving only the necessary dependencies. Thus, causal modeling enables better abstraction, helping the agent focus on the critical aspects and indirectly improving sampling efficiency. Moreover, abstraction is closely related to representation learning (Schölkopf et al., 2021). Causal representations have better robustness and transferability than correlation-based representations.

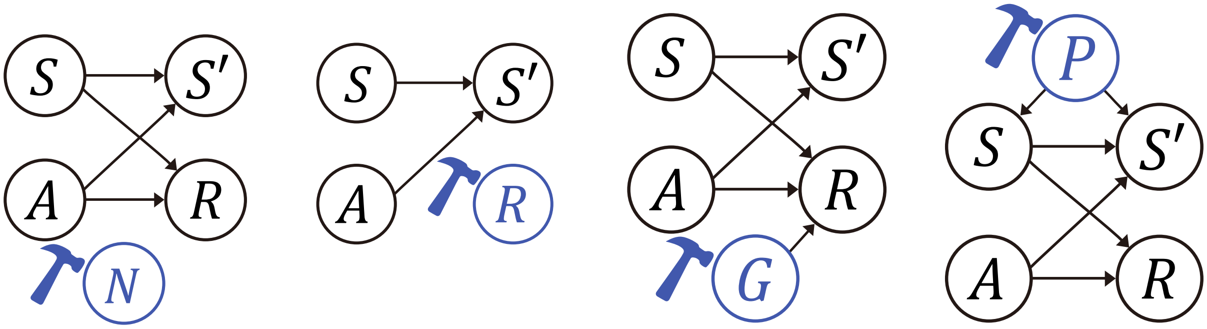
Another common idea for improving sample efficiency is to design more effective exploration methods to help agents collect data that benefit policy learning the most (Yang et al., 2022b). Some research (Pathak et al., 2017; Burda et al., 2022) has drawn on the concept of intrinsic motivation in developmental psychology (Ryan & Deci, 2000; Barto, 2013) to motivate agents to explore unknown environments by providing intrinsic rewards for exploratory behaviors. This approach provides delayed feedback to the agent, unlike approaches that directly encourage exploration while collecting data. The uncertainty-based approach follows the principle of *optimism in the face of uncertainty* (Ciosek et al., 2019; Lee et al., 2021a). It encourages agents to prioritize the exploration of areas of high epistemic uncertainty (by treating uncertainty as a bonus) in order to locate high-reward regions with fewer interactions. However, not all regions of high uncertainty are equally important. Causal modeling helps the agent detect the key regions, such as the region near the target in robotic manipulation, by narrowing down the scope.

Despite simplifying the problem and collecting informative data, sample efficiency also involves using the data more effectively. Model-based reinforcement learning (MBRL) (Wang et al., 2019; Luo et al., 2022) is a straightforward idea. As long as the learning of the environment model is accurate and efficient, the agent can use the learned model to generate data without accessing the environment. Causal models are more powerful than traditional correlation-based models because they are more robust and enable counterfactual reasoning. By explicitly considering exogenous variables, causal models can produce samples of higher quality.

3.2 Generalizability in Reinforcement Learning

3.2.1 The Issue of Generalizability in Reinforcement Learning

Generalizability in RL is another major challenge to the deployment of RL algorithms in the real world. It refers to the ability of a trained policy to perform well in new, unseen situations (Kirk et al., 2022). Training and testing in the same environment has been a notorious problem in the RL community (Irpan, 2018). While people often expect RL to work reliably in different (but similar) environments or tasks, traditional RL algorithms are typically designed to solve a single MDP. They can easily overfit the environment, failing to adapt to minor changes. Even in the same environment, RL algorithms can produce widely varying results with different random seeds (Zhang et al., 2018a;b), indicating instability and overfitting. Lanctot et al. (2017) presented an example of overfitting in multi-agent scenarios in which a well-trained RL agent struggles to adapt when the adversary changes its strategy. A similar phenomenon was observed by Raghu et al. (2018). Additionally, the real world is non-stationary and constantly changing (Hamadanian et al., 2022), so a good RL algorithm must be robust in order to handle these changes. When the situation varies, agents should be able to transfer their skills effectively rather than starting from scratch.



(a) Generalize to noise or irrelevant variables.

(b) Generalize to different reward assignments.

(c) Generalize to different goals.

(d) Generalize to different physical properties.

Figure 5: Different types of generalization problems in reinforcement learning represented by causal graphs.

3.2.2 Why Causal Modeling Helps Improve Generalization and Facilitate Knowledge Transfer?

In RL, generalization involves different samples or different environments. Kirk et al. (2022) proposed using contextual MDP (CMDP) (Hallak et al., 2015) to formalize the generalization problem in RL. CMDP is similar to a standard MDP, but it captures the variability in a set of environments or tasks determined by context. These contextual variables may impact the state space, transition function, reward function, or emission function (which maps states to observations) differently. They can represent a variety of factors, such as random seeds, goals, colors, and difficulty of game levels.

Some previous studies have shown that data augmentation can improve generalization (Lee et al., 2020; Wang et al., 2020a; Yarats et al., 2021), particularly for vision-based control. This process involves generating new data by randomly shifting, mixing, or perturbing observations, which makes the learned policy more resistant to similar changes. Another common practice is domain randomization. In sim-to-real reinforcement learning, researchers randomized the parameters of simulators to facilitate adaptation to reality (Tobin et al., 2017; Peng et al., 2018). Inspired by domain randomization, Wellmer & Kwok (2021) applied the dropout technique to the world model to create infinite dream worlds. OpenAI developed an automatic domain randomization that successfully solved the Rubik's Cube problem (OpenAI et al., 2019). Additionally, some approaches have attempted to incorporate inductive bias by designing special network structures to improve generalization performance (Kansky et al., 2017; Higgins et al., 2017; Zambaldi et al., 2019; Raileanu & Fergus, 2021).

Although these works have made significant progress in improving generalization, the connection between generalization and the underlying data generation process often remains unclear, and it is essential for identifying the factors of changes. From the causal perspective, in-distribution generalization does not involve manipulating the data generation process. As long as the SCM is accurate, it prevents the agent from overfitting unstable non-causal relationships and naturally allows the agent to generalize to data generated by the same causal model. This implies that the agent can perform well on unseen data without further training.

In the out-of-distribution (OOD) generalization scenario, the data distribution is changed, which can be interpreted as an external intervention in the data generation process. Different interventions result in different data distributions, which can pose different challenges to generalization. Figure 5 illustrates some examples of the generalization problems corresponding to different interventions. Causal modeling enables us to analyze and discuss the changes in the data distribution explicitly, distinguishing between what changes and what remains invariant. When more domain knowledge is available, we can further decompose the endogenous variables to derive a more fine-grained data generation process (this can also be achieved through causal learning). Causal modeling factorizes the joint distribution of data along the causal direction (See equation 1). In general, changes in the data distribution only relate to a small number of causal mechanisms. By virtue of the independent causal mechanisms principle (Schölkopf et al., 2021), the agent only needs to

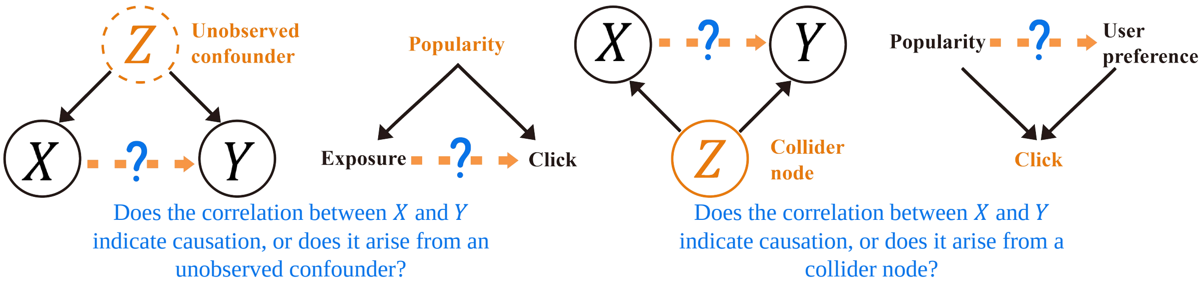

Z
Unobserved confounder
X
?
Y
Popularity
Exposure
?
Click
Does the correlation between X and Y indicate causation, or does it arise from an unobserved confounder?
X
?
Y
Z
Collider node
Popularity
?
User preference
Click
Does the correlation between X and Y indicate causation, or does it arise from a collider node?

Figure 6: Causal graphs illustrating the two types of spurious correlations, with examples from real-world applications.

fine-tune a few modules to adapt to new environments or tasks. Other modules remain unchanged and are directly reusable. In particular, when the changed module is task-independent (e.g., the background color in a robotic manipulation task), we can train a policy that only focuses on *invariance*. In that case, the agent can directly reuse the policy in any new situation, achieving zero-shot generalization.

3.3 Spurious Correlations in Reinforcement Learning

3.3.1 The Issue of Spurious Correlation in Reinforcement Learning

Learning decisions from data alone is insufficient because correlation does not imply causation. A spurious correlation is a relationship between two variables that appears to be causal but is actually caused by a third variable, bringing undesired bias to the learning problem. This phenomenon occurs widely in machine learning applications, with a few typical examples given below.

- In recommendation systems, both user behavior and preferences are influenced by conformity. If the recommender ignores conformity, it may overestimate a user's preference for certain items (Gao et al., 2022);

- In image classification, if dogs frequently appear with grass in the training set, the classifier may label an image of grass as a dog. This is because the model relies on the background (irrelevant factors) instead of the pixels corresponding to dogs (the actual cause) (Zhang et al., 2021a; Wang et al., 2021c).

- When determining the ranking of tweets, the use of gender icons in tweets is usually not causally related to the number of likes; their statistical correlation comes from the topic, as it influences both the choice of icon and the audience. Therefore, it is not appropriate to determine the ranking by gender icons (Feder et al., 2022).

If we want to apply RL in real-world scenarios, it is important to be mindful of spurious correlations, especially when the agent is working with biased data. For instance, when optimizing long-term user satisfaction in multiple-round recommendations, there is often a spurious correlation between exposure and clicks in adjacent timesteps. This is because they are both influenced by item popularity. From another perspective, when we observe a click, it may depend on user preference or item popularity, which creates a spurious correlation between the two factors. In both scenarios, if the agent ignores causality, it will make incorrect predictions or decisions, such as by creating filter bubbles by only recommending popular items (a suboptimal policy for both the system and the user). In a nutshell, if the agent learns a spurious correlation between two variables, it may mistakenly believe that changing one will affect the other, even though there is no causal relationship between them in the underlying data generation process. This misunderstanding can lead to suboptimal or even harmful behavior.

3.3.2 Why Causal Modeling Helps Address Spurious Correlations?

From the causal perspective, spurious correlations arise when the data generation process involves unobserved confounders (common cause) or when a collider node (common effect) serves as the condition. The former leads to confounding bias, while the latter results in selection bias. See Figure 6 for a visual interpretation of these phenomena. With causal graphs, we can trace the source of spurious correlations by closely scrutinizing the data generation process. In order to eliminate bias, it is necessary to harness causality instead of relying on statistical correlations. This is where causal reasoning comes in: It provides tools that allow us to analyze and deal with confounding and selective bias (Pearl, 2009b; Glymour et al., 2016), helping RL agents estimate the causal effects of decision-making problems more accurately. Therefore, causal modeling helps avoid misinterpretation of the environment or task that leads to suboptimal policies. Using causal graphs and causal reasoning enables RL agents to improve the correctness of inferences in decision-making.

3.4 Considerations Beyond Return

In general, RL focuses on maximizing returns. However, as automated decision systems based on RL become a more prevalent feature of our daily lives, it is crucial to pay attention to how RL agents interact with people and how they may influence society.

3.4.1 Explainability in Reinforcement Learning

Explainability in RL refers to the ability to understand and interpret the decisions of an RL agent. It is important to both researchers and general users. Explanations reflect the knowledge learned by the agent, facilitate in-depth understanding, and allow researchers to participate efficiently in the design and continual optimization of an algorithm. In addition, explanations provide the internal logic of the decision-making process. When agents outperform humans, we can extract knowledge from the explanations to guide human practice in a specific domain. For general users, the explanation presents the reasons for the decision, thus deepening the user's understanding of intelligent agents and increasing the user's confidence in the agent.

3.4.2 Achieving Explainability through Causal Modeling

Explainable RL methods can be divided into two categories: post hoc and intrinsic approaches (Puiutta & Veith, 2020; Heuillet et al., 2021). The former provides explanations after execution, while the latter is inherently transparent. Post-hoc explanations are generally established based on correlations, such as the saliency map approach (Greydanus et al., 2018; Mott et al., 2019). As we mentioned earlier, conclusions based on correlations can be unreliable, failing to answer causal questions. On the other hand, intrinsic explanations can be achieved using easy-to-understand algorithms, such as linear regression or decision trees (Coppens et al., 2019). However, the limited model capability may be insufficient for explaining complicated behaviors (Puiutta & Veith, 2020).

Humans possess an innate and powerful ability to establish connections between different events through a "mental causal model (Sloman, 2005)". We frequently use causal language, such as "because," "therefore," and "if only," in our daily lives to facilitate communication and collaboration. Using causal models allows for natural and flexible explanations, as it does not require selecting specific algorithms or models. In RL, causality-based explainability provides stable support for the agent's decisions and helps us understand how the agent interprets the environment and task. When the agent makes mistakes, we can respond with more tailored solutions.

3.4.3 Fairness in Reinforcement Learning

As machine learning becomes more prevalent in our daily lives, stakeholders, such as business owners, general users, and policymakers are realizing the importance of fairness. This concept applies to any type of automated system and decision support system, including those based on RL. In particular, RL agents should strive to genuinely benefit people and promote social good rather than causing discrimination or harm against specific individuals or groups. In addition, fairness issues in the real world are often dynamic (Gajane et al., 2022), involving multiple decisions. For example, a hiring process is typically a sequential decision process, and the actions may have cumulative effects on fairness. Ignoring the dynamic nature of a system may lead to unintended unfairness (Liu et al., 2018; Creager et al., 2020; D'Amour et al., 2020).

3.4.4 Achieving Fairness through Causal Modeling

In legal cases, fairness or discrimination often involves a counterfactual statement (Pearl & Mackenzie, 2018). For example, in *Carson v. Bethlehem Steel Corporation* (1996) [2], the ruling made it clear that determining whether an individual or group would be treated differently if they changed only the sensitive attribute (e.g., sex, age, and race) while holding other factors constant is at the heart of discrimination issues. In machine learning, researchers often use demographic parity (Zafar et al., 2015; Wen et al., 2021), a metric built on the difference between the conditional distributions of different groups, to study fairness. However, as pointed out by Kusner et al. (2017) and Zhang & Bareinboim (2018), correlation-based metrics ignore the causal relationships behind the data generation process, resulting in an inaccurate measurement of fairness, which may increase discrimination in certain situations. As illustrated in Figure 3, conditional probabilities and counterfactual probabilities are significantly different. The SCM framework makes it clear that fairness issues require a comparison of the differences in causal effects between the factual and imagined worlds. Therefore, using causal models is a principled way of studying fair reinforcement learning, which offers a new perspective compared to non-causal approaches.

3.4.5 Safety in Reinforcement Learning

Safety is a crucial aspect of RL (Garcıa & Fernández, 2015; Gu et al., 2022). RL agents may sometimes exhibit unexpected or unpredictable behavior, particularly when encountering new or unforeseen situations. This issue can pose a significant risk in safety-critical applications, such as healthcare or autonomous vehicles, where a single mistake could result in severe consequences. Additionally, RL agents may prioritize higher returns over their own safety, known as the agent safety problem (Fulton & Platzer, 2018; Beard & Baheri, 2022). A typical example is robotic control, where agents sacrifice their lifespan for a higher mission completion rate. In short, ensuring safety in RL is vital for preventing accidents or other harmful events in reality.

3.4.6 Achieving Safety through Causal Modeling

Researchers often use constrained MDPs (Altman, 1995; 1999) to model the safety issues in RL. Constrained MDPs extend MDPs by incorporating a constraint set that represents safety concerns. Accordingly, most relevant studies have focused on solving constrained optimization problems (Achiam et al., 2017; Chow et al., 2017), rarely considering causality. With the aid of causal models, we can formalize prior knowledge more effectively and obtain valuable insights, such as explanations, by analyzing the generation processes of unsafe states or actions. When RL agents violate safety constraints, causal models can help researchers and experts better understand the causes of unexpected outcomes and design solutions to prevent them from happening again. Additionally, causal models can be utilized for counterfactual policy evaluation, allowing for testing and identifying potential security issues before deploying an RL agent in real-world applications. Overall, causal modeling helps ensure RL techniques are used safely and responsibly, avoiding catastrophic consequences.

To summarize, we discussed several key challenges of RL in this section and considered why causal modeling plays a crucial role in solving or mitigating these challenges. Next, we review recent advances in causal RL.

4 Existing Work Relating to Causal Reinforcement Learning

In the previous section, we highlighted the significance of causal RL. However, there is a lack of clarity and coherence in the existing literature on this topic, primarily because causal modeling is more of a mindset than a specific issue. It provides principles and insights for solving a wide range of problems. This section reviews existing approaches to causal RL that address the four critical challenges outlined in section 3. We

[2] `https://caselaw.findlaw.com/us-7th-circuit/1304532.html`

Table 1: Selected methods that employ causal modeling to optimize sample efficiency.

Category	Paper	Technique	Environments or Tasks	Source
Representaion Learning	Sontakke et al. (2021)	Causal representation learning	CausalWorld	ICML
	Lee et al. (2021b)	Intervention Domain randomization	Manipulation (Isaac Gym)	ICRA
	Huang et al. (2022b)	Causal dynamics learning	CarRacing VizDoom	ICML
	Wang et al. (2022)	Causal dynamics learning	Chemical Manipulation (robosuite)	ICML
Directed Exploration	Seitzer et al. (2021)	Intervention	Manipulation (OpenAI)	NeurIPS
Data Augmentation	Buesing et al. (2019)	Counterfactual reasoning	Sokoban	ICLR
	Lu et al. (2020)	Counterfactual reasoning	Cartpole MIMIC-3	NeurIPS Workshop
	Pitis et al. (2020)	Counterfactual reasoning	Spriteworld Pong (Roboschool) Manipulation (OpenAI)	NeurIPS
	Zhu et al. (2021)	Counterfactual reasoning	CausalWorld	OpenReview

organize these approaches based on their problem settings and solution methods with the goal of better understanding their connections and relationships.

4.1 Causal Reinforcement Learning for Addressing Sample Inefficiency

Causal modeling offers some useful principles for designing sample-efficient RL algorithms. We can organize these principles into three lines of research: representation learning, directed exploration, and data augmentation. The representative works are shown in Table 1.

4.1.1 Representation Learning for Sample Efficiency

A good representation of the environment can be beneficial for sample-efficient RL. By providing a compact and informative representation of the environment, an RL agent can learn more effectively with fewer samples. This is because a good representation can help the agent identify important features of the environment and abstract away unnecessary details, allowing the agent to learn more generalizable policies and make better use of its experiences.

The research of Sontakke et al. (2021) involved clustering the trajectories generated from various environments with different physical properties and using the clustering outcomes as a causal representation of the environment. Using the state augmented by the causal representation, the learn policies exhibit outstanding zero-shot generalization ability, and require only a small number of training samples to converge in new environments. Causal modeling also inspires state abstraction. Lee et al. (2021b) used interventions to identify the state variables that are important for successful task completion, reducing the dimensionality of the state space and simplifying the problem. Another approach to state abstraction is causal dynamics learning, which identifies the causal relationships between different variables by learning a causal dynamics model. Huang et al. (2022b) proposed using action-sufficient state representations, a minimal set of state variables containing sufficient decision-making information, to improve sample efficiency in RL. Wang et al. (2022), on the other hand, studied task-independent state abstraction. Unlike previous works, they adopted a shared structured dynamical model that removes irrelevant dependencies while preserving relevant state variables that may be used in new tasks. To conduct planning or model-based RL in downstream tasks, an agent only needs to learn a reward predictor.

4.1.2 Directed Exploration for Sample Efficiency

While a good representation of the environment is beneficial, it is not necessarily sufficient for sample efficient RL (Du et al., 2020). To improve sample efficiency, researchers have been studying directed exploration, strategies that guide the agent to explore specific parts of the state space that are believed to be more informative or more likely to yield high rewards. This can be done by giving bonuses to exploratory behaviors that discover novel or uncertain states. From a causal perspective, not all regions of high uncertainty are

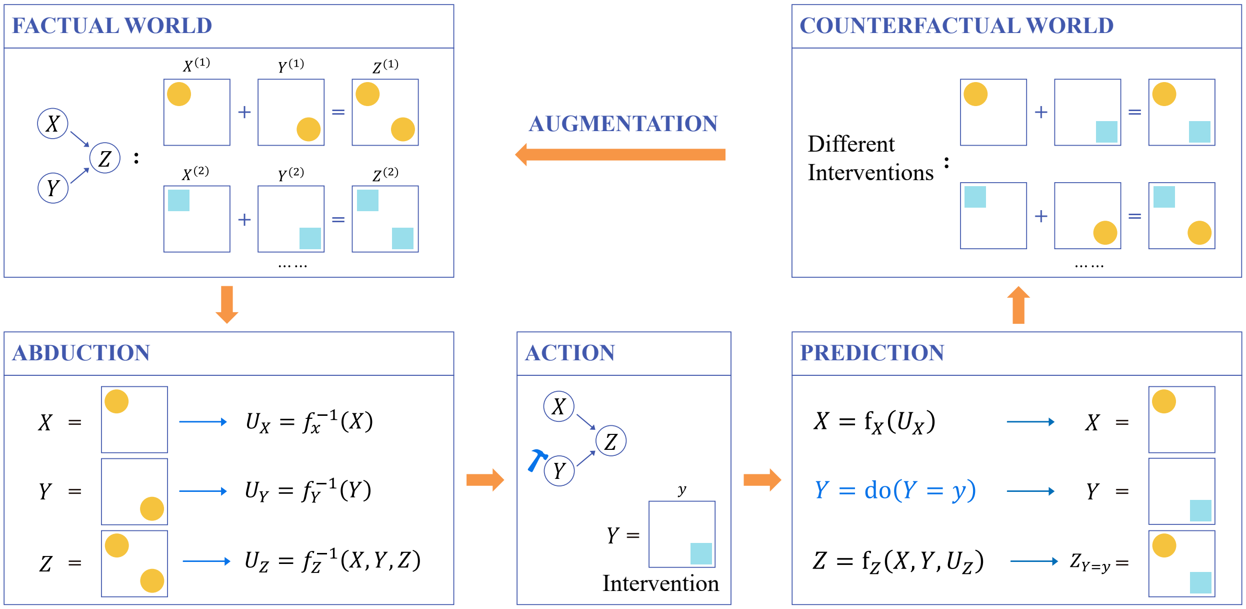


Figure 7: An example of counterfactual data augmentation following the counterfactual reasoning procedure: abduction, action, and prediction. The outcome of this procedure is then used to augment the training data observed in the factual world.

equally important. Only those that form a causal relationship with the success of the task are worth exploring. As an example, Seitzer et al. (2021) studied the problem of directed exploration in robotic manipulation tasks. Valuable data can be generated only if the agent touches the target object, a prerequisite for learning complex manipulation tasks. The authors proposed a method of measuring the effect of an action on the object (a causal quantity) and incorporating it into exploration, greatly improving the sample efficiency of robotic manipulation tasks. Furthermore, an agent's curiosity about causality can drive it to acquire useful behaviors and knowledge even without extrinsic rewards. Inspired by the principle of independent causal mechanisms, Sontakke et al. (2021) proposed a method to formalize causal curiosity, allowing the agent to perform experiments that facilitate understanding of the environment. Experiments show that RL agents that are pre-trained with causal curiosity can learn to solve new tasks much faster.

4.1.3 Data Augmentation for Sample Efficiency

Data augmentation is a common machine learning technique that aims to improve the performance of algorithms by generating additional training data. Counterfactual data augmentation is a causality-based approach that uses a causal model to imitate the environment and generate data that is not observed in the real world. This is particularly useful for RL problems because collecting large amounts of real-world data is often difficult or expensive. By simulating different counterfactual scenarios, RL agents can determine the effects of different actions without interacting with the environment, leading to more sample-efficient learning.

The implementation of counterfactual data augmentation follows a counterfactual reasoning procedure that consists of three steps (Pearl, 2009a), as demonstrated in Figure 7:

1. **Abduction** is about using observed data to infer the values of the exogenous variables $\mathcal{U}$;
2. **Action** involves modifying the structural equations of the variables of interest in the SCM; and
3. **Prediction** uses the modified SCM to generate counterfactual data by plugging the exogenous variables back into the equations for computation.

Table 2: Selected methods that employ causal modeling to improve generalizability.

Category	Paper	Technique	Environments or Tasks	Source
Irrelevant variables	Zhang et al. (2020a)	Causal representation learning	Toy [3] dm_control	ICML
	Bica et al. (2021b)	Causal representation learning	OpenAI Gym MIMIC III	NeurIPS
	Wang et al. (2022)	Causal dynamics learning	Chemical Manipulation (robosuite)	ICML
	Saengkyongam et al. (2022)	Causal representation learning	Toy	ICML Workshop
	Ding et al. (2022)	Causal discovery	Manipulation Unlock (Mini-grid) Crash (highway-env)	NeurIPS
Dynamics	Sontakke et al. (2021)	Causal representation learning	CausalWolrd	ICML
	Lee et al. (2021b)	Intervention Domain randomization	Manipulation (Isaac Gym)	ICRA
	Zhu et al. (2021)	Counterfactual reasoning	CausalWolrd	OpenReview
	Guo et al. (2022)	Mediation analysis	Gym-MuJoCo	ICLR
Tasks	Eghbal-zadeh et al. (2021)	Causal representation learning	Grid-world	ICLR Workshop
	Pitis et al. (2022)	Counterfactual reasoning	Spriteworld Pong (Roboschool) Manipulation (OpenAI)	ICML Workshop
Mixed	Zhang & Bareinboim (2017)	Causal reasoning	Toy	IJCAI
	Dasgupta et al. (2018)	Meta learning	Toy	ICLR Workshop
	Nair et al. (2019)	Causal discovery	Light switch control	arXiv
	Huang et al. (2022a)	Causal dynamics learning	Cartpole Pong (Atari)	ICLR
Others	Zhu et al. (2022b)	Causal discovery Causal dynamics learning	Toy Gym-MuJoCo	arXiv

While MBRL methods can also generate samples with the learned models, they lack the means to model exogenous variables. This can result in underfitting when the distribution of exogenous variables is complex (Buesing et al., 2019). In contrast, counterfactual data augmentation explicitly considers exogenous variables using the SCM framework and is able to generate higher-quality samples. From a Bayesian perspective, traditional MBRL approaches use a fixed prior distribution for exogenous variables, whereas counterfactual data augmentation uses observed data to estimate the posterior distribution.

Buesing et al. (2019) proposed counterfactually-guided policy search (CF-GPS), an algorithm for learning policies in POMDPs that uses SCMs to evaluate counterfactual actions. The CF-GPS method improves traditional model-based RL algorithms by inferring the posterior distribution of exogenous variables and considering alternative outcomes. Lu et al. (2020) proposed a sample-efficient RL algorithm to address the problems of mechanism heterogeneity and data scarcity. SCM is used to model the environment, and counterfactual reasoning is performed to evaluate the potential consequences of alternate possibilities. This method avoids actual exploration and mitigates the bias that arises from limited experience. Pitis et al. (2020) proposed to generate counterfactual data by swapping factorized subprocesses of observed trajectory pairs. They introduced local causal models and showed how they could simplify counterfactual reasoning. Their experiments demonstrated that the proposed counterfactual data augmentation method significantly improves sample efficiency in various RL settings. Zhu et al. (2021) studied the application of counterfactual reasoning in robotic manipulation tasks, which increases the diversity of generated rollouts. The proposed method uses SCM to model the underlying dynamics and can generate samples with rarely seen objects to improve sample efficiency.

4.2 Causal Reinforcement Learning for Addressing Generalizability

Decision-making problems in the real world are ever-changing and hard to predict. Therefore, RL algorithms must be able to perform well in new and unseen situations during deployment, also known as generalizability.

[3] The term “toy” refers to simple, synthetically constructed datasets or simulation environments that are used to experimentally verify findings. It is not a concrete environment or task. We use this term consistently throughout the paper.

Generalization involves many types of problems. Zero-shot generalization requires the agent to only learn in training environments and be tested on the unseen. While this setting is appealing, it is sometimes infeasible in practice. Alternatively, adaptation (Zhang et al., 2015; Gong et al., 2016) assumes the agent can receive additional training in test domains, which encompasses a variety of settings such as transfer RL (Zhu et al., 2020), multitask RL (Vithayathil Varghese & Mahmoud, 2020), or lifelong RL (Khetarpal et al., 2022). A great deal of research has considered generalization in RL (Kirk et al., 2022), but there is still a lack of understanding of what capabilities an agent needs to achieve generalization and what generalization can be expected from a learning algorithm. Causal models provide a possible way of answering these questions. This section classifies existing causal RL algorithms according to their specific generalization goals. The representative works are shown in Table 2.

4.2.1 Generalize to Different Environments

First, we consider how to generalize to different environments. From a causal perspective, different environments share most of the causal mechanisms but differ in certain modules, resulting from different interventions in the state variables. Based on the causal relationship between these variables, we can further divide existing works into two categories: generalizing to irrelevant variables and generalizing to different dynamics.

To enhance the ability to generalize to irrelevant factors, RL agents must master the invariance in the data generation process. Zhang et al. (2020a) studied the problem of generalizing to different observation spaces within the block MDP framework. This problem is common in reality, such as when robots are equipped with different types of cameras and sensors. Within the block MDP framework, the observation may be infinite, but it can uniquely determine the state of the environment (finite but unobservable) and maintain the Markovian property. The authors proposed using invariant prediction to learn the causal representation of the reward, which allows the agent to achieve zero-shot generalization in a new observation space. Bica et al. (2021b) introduced invariant causal imitation learning (ICIL) that allows an imitation policy to be learned from multiple domains and then deployed in new environments. The ICIL approach achieves this goal by learning a shared representation of causal variables that are consistent across domains. Wang et al. (2022) studied the causal dynamics learning problem, which attempts to eliminate irrelevant variables and unnecessary dependencies between actions and state variables. Saengkyongam et al. (2022). connected causality, invariance, and contextual bandits. They introduced the concept of policy invariance and showed that an optimal invariant policy would generalize across environments in the presence of unobserved variables. Ding et al. (2022) proposed a novel solution to the generalization problem in goal-conditioned reinforcement learning (GCRL) by treating the causal graph as a latent variable and optimizing it using a variational likelihood maximization approach. This method trains agents to discover causal relationships and learn a causality-aware policy, which is robust against irrelevant variables.

Generalizing to new dynamics is a broader issue. It may involve variations in physical properties (e.g., gravitational acceleration, as shown in Figure 5d), differences between the simulation environment and reality, changes in the range of attribute values, etc. Sontakke et al. (2021) proposed training RL agents to categorize and infer causal factors in the environment, which would be accomplished through an innovative intrinsic motivation called causal curiosity. The agents can learn semantically meaningful behaviors in a self-supervised manner and the learned causal representations empower them to generalize to unseen contexts. Lee et al. (2021b) studied how to teach robots to perform manipulation tasks by conducting interventions, which helps determine the relevant relationship between the state and action spaces. The robot exhibits excellent sim-to-real generalizability after training with domain randomization on the relevant features. Zhu et al. (2021) developed an algorithm to improve the ability of agents to generalize to rarely seen or unseen object properties. This algorithm uses SCMs to model the environmental dynamics, allowing the agent to reason about what would have happened if the object had a different attribute value, which leads to improved generalizability. Guo et al. (2022) investigated the unsupervised dynamics generalization problem, which allows the model to generalize to new environments. The authors followed the intuition that data from the same trajectory/similar environments should have similar properties (hidden variables) that lead to similar causal effects. They used conditional direct effects in mediation analysis to measure similarity. The experimental results show that the learned model performs well in new dynamics.

4.2.2 Generalize to Different Tasks

Another important topic is how to generalize to different tasks. In the SCM framework, different tasks are created by altering the structural equation of the reward variable or its parent nodes on the causal graph. These tasks have the same underlying environmental dynamics, but the rewards are assigned differently.

Eghbal-zadeh et al. (2021) introduced causal contextual RL, in which the agent should learn adaptive policies that can adapt to new tasks specified by context variables. They proposed a contextual attention module that allows agents to incorporate disentangled features as contextual factors, achieving better generalization than non-causal agents. In order to make RL more effective in complex, multi-object environments, Pitis et al. (2022) proposed that local factors in transition dynamics should be recognized and used. They proposed a new framework called model-based counterfactual data augmentation that leverages the local structure to generate counterfactual transitions, allowing the model to generalize to OOD tasks.

Furthermore, in reality, generalization may involve changes in both the environmental dynamics and the task. Several studies have explored this problem from a causal viewpoint. Zhang & Bareinboim (2017) used causal inference to improve knowledge transfer in reinforcement learning. Their method addresses the issue of transferring knowledge between bandit agents when standard techniques fail to identify causal effects. They introduce a new identification strategy that involves deriving bounds on the distribution of arms using structural knowledge and transferring these bounds in new bandit algorithms. Dasgupta et al. (2018) investigated whether meta-RL can teach agents to perform causal reasoning. The experimental results showed that the agents successfully learned to conduct interventions, which supports subsequent tasks that reward accurate causal inferences. The agent also learns to make sophisticated counterfactual predictions, and this emergent ability can effectively generalize to new causal structures. Nair et al. (2019) developed an approach using directed acyclic graphs to impart causal knowledge to the agent. They utilized attention mechanisms to allow the agent to generate a causal graph based on its visual observations and use it to make informed decisions. The experiments demonstrated that the agent effectively generalizes to new tasks and environments with unknown causal structures. Huang et al. (2022c) proposed AdaRL, a framework for adaptive RL that allows for fast adaptations to new environments, tasks, or observations. They used a compact graphical representation to encode structural relationships between variables in decision-making problems, which allows for efficient policy adaptation to new domains with only a few samples.

4.2.3 Other Generalization Problems

In offline RL, the agent can only learn from pre-collected datasets. With this setting, agents may encounter previously unseen state-action pairs during the testing phase, leading to the distributional shift issue (Levine et al., 2020), a common challenge in offline RL. Most existing approaches mitigate this issue through conservative or pessimistic learning (Fujimoto et al., 2019; Kumar et al., 2020; Yang et al., 2021b), rarely considering the combination of offline RL and generalization (Kirk et al., 2022). Zhu et al. (2022b) proposed a solution to generalize to unseen states. They recovered the causal structure from offline data by using causal discovery techniques. The experimental results suggested that the causal world model exhibits better generalization performance than a traditional world model.

4.3 Causal Reinforcement Learning for Addressing Spurious Correlations

As we stated in section 3.3, an RL agent is vulnerable to spurious correlations during the process of understanding the environment and task. Depending on the causal structure behind the decision-making problem, the spurious correlation can be one of two types: one corresponding to confounding bias caused by the fork structure and the other corresponding to selective bias caused by the collider structure (Figure 6). We divide the existing methods into two categories accordingly. In particular, we also include work on imitation learning (IL) and off-policy evaluation (OPE), as they are both closely related to policy learning in RL. The representative works are shown in Table 3.

[4] Causal graph as a technique refers to using causal graphs to describe the data generation process, and designing graphical criteria for determining properties such as identifiability or developing algorithms based on causal graphs.

Table 3: Selected methods that employ causal modeling to address spurious correlations.

Category	Paper	Technique	Environments or Tasks	Source
Confounding bias - MAB	Forney et al. (2017)	Counterfactual data fusion	Toy	ICML
Confounding bias - MDP: IL	Zhang et al. (2020b)	Causal graph [4]	Toy	NeurIPS
	Kumor et al. (2021)	Causal graph	Toy	NeurIPS
	Swamy et al. (2022)	Instrumental variable regression	LunarLander PyBullet Gym	ICML
Confounding bias - MDP: OPE	Namkoong et al. (2020)	Sensitivity analysis	Toy	NeurIPS
	Bennett et al. (2021)	Proximal causal inference	Toy GridWorld	AISTATS
Confounding bias - MDP	Lu & Lobato (2018)	Backdoor adjustment	Pendulum Cartpole MNIST	arXiv
	Rezende et al. (2020)	Front-door adjustment Backdoor adjustment	Toy MiniPacman DeepMind Lab	arXiv
	Wang et al. (2021b)	Front-door adjustment Backdoor adjustment	-	NeurIPS
	Liao et al. (2021)	Instrumental variable regression	-	arXiv
	Gasse et al. (2021)	Do-calculus	Toy	arXiv
	Yang et al. (2022a)	Causal graph	Toy Cartpole LunarLander	AAAI
Selection bias	Bai et al. (2021)	Inverse probability weighting	Manipulation (OpenAI)	TCYB
	Deng et al. (2021)	Causal graph	D4RL	arXiv

4.3.1 Addressing Confounding Bias

We start by introducing several techniques that eliminate confounding bias using causal reasoning (Glymour et al., 2016). One of the most common is the backdoor adjustment approach, a technique used to accurately estimate the causal effect between treatment and outcome variables by controlling for variables that satisfy the backdoor criterion. The backdoor variables (e.g., the confounders) can block all spurious paths and ensure that causality fully accounts for the correlation between the treatment and outcome variables. If it is impossible to find a set of covariates that meets the backdoor criterion (e.g., all variables on the backdoor path are unobservable), one may use the front-door adjustment method. This approach involves estimating the causal effect by controlling for variables on the directed path from the treatment variable to the outcome variable. When neither method is feasible, one can use do-calculus (Pearl, 1995), a complete axiomatic system that retrieves the covariates that can help to identify the causal effect or reports failure whenever the effect is unidentifiable. This is not an exhaustive list of techniques for addressing confounding bias, and many other methods can be used to mitigate this issue.

The MAB problem can be thought of as a single-step decision-making problem with no state transitions. Forney et al. (2017) investigated a variant of this problem that involves unmeasured variables (unobserved confounders) that affect both actions and rewards. They found that counterfactual-based decision-making helps address this problem and facilitates a fusion of observational and experimental data. Zhang et al. (2020b) studied single-step imitation learning using a combination of demonstration data and qualitative knowledge about the data-generating process. They proposed a graphical criterion for determining the feasibility of imitation in the presence of unobserved confounders and a practical procedure for estimating valid imitating policies with confounded expert data. This approach was then extended to the sequential setting in a subsequent paper (Kumor et al., 2021). Swamy et al. (2022) designed an algorithm for imitation learning with corrupted data. They proposed to use instrumental variable regression (Stock & Trebbi, 2003), a well-known causal inference technique, to break up spurious correlations.

Several research focused on the off-policy evaluation (OPE) problem, which can estimate the performance of policies before they are deployed. For example, Namkoong et al. (2020) assessed the robustness of OPE methods under unobserved confounding by developing worst-case bounds on the performance of an evaluation policy. They also proposed an efficient procedure for computing the worst-case bounds, allowing for a reliable selection of policies. Bennett et al. (2021), on the other hand, proposed a new estimator for OPE in infinite-horizon RL with unobserved confounders. They focused on specific settings in which the model and the policy value are identifiable.

Lu & Lobato (2018) studied the connection between causal inference and RL. They presented a method called deconfounding RL, which allows for learning good policies from historical data affected by unobserved factors. This method is superior to traditional RL methods when applied to observational data with confounders. Liao et al. (2021) also focused on the offline setting. They found that unobserved confounders usually influence the actions in observational studies. They proposed an algorithm that helps efficiently identify the transition dynamics in RL using instrumental variables. Wang et al. (2021b), on the other hand, proposed a method for incorporating offline data in an online setting, accounting for confounding variables that may affect the data's accuracy. This method effectively adjusts for confounding biases and achieves a smaller regret than the optimal online-only approach. Gasse et al. (2021) studied a model-based method that combines interventional and observational data to learn a latent-based causal transition model, which is then used to solve the POMDP problem via deconfounding. Yang et al. (2022a) proposed an algorithm called causal inference Q-network to deal with the confounding bias raised by multiple types of observational interferences. Due to the advantages of causal inference on deconfounding, this method is more resilient to interferences, leading to better performance. Rezende et al. (2020) discussed the use of partial models in RL, a model-based approach that does not require modeling the full (and usually high-dimensional) observation. Partial models can be problematic when they are confounded by the parts they do not model, leading to incorrect planning. The authors proposed a solution to overcome the deficiency by ensuring the partial models are causally accurate.

4.3.2 Addressing Selection Bias

Selective bias occurs when data samples fail to represent the target population correctly. For example, selective bias arises when researchers seek to understand the effect of a certain drug on curing a disease by investigating patients in a selected hospital. This is because those patients may differ significantly from the population regarding where they reside, their social status, and their wealth, making them unrepresentative.

Bai et al. (2021) investigated the selective bias associated with using hindsight experience replay (HER) in goal-conditioned reinforcement learning (GCRL) problems. In particular, HER relabels the goal of each collected trajectory, allowing the agent to learn from failures (not reaching the original goal). The concern is that the relabeled goal distribution cannot properly represent the original goal distribution; thus, the agent trained with HER is biased. The authors proposed to use the inverse probability weighting technique in causal inference for learning, which allows the agent to improve sample efficiency with the help of HER while avoiding the bias that arises due from the relabeling, achieving promising results on a range of robotic manipulation tasks. Deng et al. (2021) viewed the offline RL problem through the lens of selective bias. An agent is vulnerable to the spurious correlation between uncertainty and decision-making in the offline setting, being prone to learning suboptimal policies. With causal modeling, we can see that the empirical return is the outcome of both uncertainty and actual return. Since it is impossible to eliminate uncertainty by acquiring more data in the offline setting, a causal-unaware agent may mistakenly believe that there is a causal relationship between uncertainty and return. As a result, it prefers policies that achieve high returns by chance (high uncertainty). The authors propose quantifying uncertainty and using it as a penalty term in the learning process. The results showed that this method outperforms various baselines that do not consider causality in the offline setting.

4.4 Beyond Return with causal Reinforcement Learning

As automated decision systems based on RL become widely used in various activities of human society, we will have various concerns about these systems. In this section, we discuss how to address these concerns through causal modeling. The representative works are shown in Table 4.

Table 4: Selected methods that employ causal modeling for goals beyond maximizing returns.

Category	Paper	Technique	Environments or Tasks	Source
Explainability	Foerster et al. (2018)	Counterfactual	StarCraft	AAAI
	Madumal et al. (2020)	Counterfactual reasoning	Gym StarCraft	AAAI
	Bica et al. (2021a)	Counterfactual reasoning	Toy MIMIC-III	ICLR
	Mesnard et al. (2021)	Counterfactual reasoning	Bandit Key-to-Door Interleaving	ICML
	Tsirtsis et al. (2021)	Counterfactual reasoning	Toy Clinical trial	NeurIPS
	Triantafyllou et al. (2022)	Counterfactual reasoning	Goofspiel	AAAI
	Herlau & Larsen (2022)	Mediation analysis	Toy DoorKey	AAAI
Fairness	Zhang & Bareinboim (2018)	Counterfactual reasoning Mediation analysis	Toy	AAAI
	Huang et al. (2022c)	Causal graph Causal reasoning	Toy	AAAI
	Balakrishnan et al. (2022)	Causal graph Counterfactual reasoning	Toy	AAAI
Safety	Hart & Knoll (2020)	Counterfactual reasoning	BARK-ML	IROC
	Everitt et al. (2021)	Causal graph	-	Synthese

4.4.1 Explainable Reinforcement Learning via Causal Modeling

A good RL agent should be capable of successfully solving tasks as well as explaining their behavior. Large language models provide excellent examples of causality enhancing communication. Humans use causal language in everyday life, so a large language model trained on large-scale human-written text can naturally and fluently communicate with humans. In general, causal RL algorithms that model the data generation process using the SCM framework are inherently explainable. We may provide prior knowledge (such as causal graphs) to the agent before training, ensuring that the agent and humans have the same understanding of the environment. With causal modeling, the agent learns causality rather than correlations. As a result, we can understand the agent's decisions at the causal level rather than relying on ambiguous correlations.

In practice, we often desire granular explanations that involve counterfactuals. The term "counterfactual" is popular in multi-agent reinforcement learning (MARL). For example, Foerster et al. (2018) proposed a method named counterfactual multi-agent policy gradients for efficiently learning decentralized policies in cooperative multi-agent systems. More precisely, counterfactual solves the challenge of multi-agent credit assignment so that agents and humans can better understand the contribution of individual behavior to the team. Some subsequent studies followed the same idea (Su et al., 2020; Zhou et al., 2022). These approaches did not perform the complete counterfactual reasoning procedure as shown in Figure 7, missing the critical step of abduction, which offers opportunities for further enhancements. More recently, Triantafyllou et al. (2022) built a connection between Dec-POMDPs and SCM, allowing them to investigate the credit assignment problem in MARL using causal language. However, this work did not aim to improve the performance of RL agents; rather, its goal was to introduce a formal definition of actual causality to establish an RL framework that supports accountability, which is key in accountable decision-making. Mesnard et al. (2021), on the other hand, studied the temporal credit assignment problem, which helps understand the influence of a particular action on future rewards.

Madumal et al. (2020) used theories from cognitive science to explain how humans understand the world through causal relationships and how these relationships can help us understand and explain the behavior of RL agents. They presented an approach that learns a SCM during RL and used this model to generate explanations of behavior based on counterfactual analysis. They conducted a study with 120 participants. The results showed that the causality-based explanations performed better in understanding, explanation satisfaction, and trust than other explanation models. Bica et al. (2021a) discussed a method for understanding and explaining expert decision processes by integrating counterfactual reasoning into batch inverse RL. By using counterfactuals, it is possible to explain expert behavior and evaluate policies in situations in which active experimentation is prohibited. Tsirtsis et al. (2021) investigated how to find the optimal counterfactual explanation for a sequential decision process. They formalized this problem as a constrained search problem, i.e., how to search for another sequence of actions that differs from the observed sequence of actions by a specified number of actions. Herlau & Larsen (2022) studied the mediation analysis technique in RL. In particular, they forced the RL agent to maximize natural indirect effects during the learning process, which allows the agent to identify key events in the decision problem, such as obtaining a key before opening a door. Using mediation analysis, the agent can learn a parsimonious descriptive causal model, which provides a new perspective on explainable RL.

4.4.2 Fair Reinforcement Learning via Causal Modeling

In addition to explainability, we also want RL agents to align with human values and avoid potential harm to human society. Fairness is an important consideration. Nevertheless, little work has been done to study fairness in RL. One straightforward idea is to quantify fairness using a statistical measure and treat it as a constraint (Balakrishnan et al., 2022). Policy optimization then becomes a constrained optimization problem that can be solved using powerful solvers such as Gurobi [5]. However, fairness poses a counterfactual problem: Would the results be different if the sensitive attribute had a different value? To accurately assess fairness, one must evaluate counterfactual quantities. Zhang & Bareinboim (2018) first introduced the SCM framework to formulate the concept of fairness, which allows researchers to assess counterfactual fairness quantitatively. Using counterfactual statements, researchers can systematically analyze different types of discrimination and their effects on decision-making. Huang et al. (2022c) studied fairness in recommendation scenarios, focusing on the bandit setting (single-step decision-making), in which sensitive attributes should not influence rewards. Liu et al. (2020) also studied the fairness problem in RL from a causal perspective. They adopted causal graphs to provide a formal analysis of the fairness problem and evaluate fairness using counterfactuals. The experimental results showed that the proposed approach could learn fair policies under multiple problem settings.

4.4.3 Safe Reinforcement Learning via Causal Modeling

Finally, safety is a fundamental challenge to making RL widely applicable in the real world. Causal modeling provides some valuable tools for studying safety. As an example, Hart & Knoll (2020) investigated the safety issue in relation to autonomous driving. Researchers can conduct counterfactual policy evaluations before deploying policies to the real world by utilizing counterfactual reasoning. The experimental results showed that their method demonstrated a high success rate while significantly reducing the collision rate. On the other hand, Everitt et al. (2021) studied the issue of reward tampering in sufficiently capable RL agents. These agents may find shortcuts to obtain rewards instead of performing the expected behavior, posing a potential safety risk. The authors presented a precise and intuitive formalization of this problem using causal diagrams and proposed design principles to prevent reward tampering.

5 Open Problems and Future Directions

In this section, we consider some significant yet underexplored topics of causal RL.

5.1 Causal Learning in Reinforcement Learning

In the previous section, we explained how causality dynamics learning - a class of methods closely related to MBRL - can improve sample efficiency and generalizability (Wang et al., 2022; Huang et al., 2022b). These methods focus on understanding the cause-and-effect relationships between variables and the process that generates these variables. Instead of using complex, redundant connections, to model the data generation process, these methods prefer a sparse, modular style. As a result, they are more efficient and stable than

[5] https://www.gurobi.com/

Table 5: The three components of causal learning.

	Available information	**Targets to identify**	**Typical questions**
Causal feature extraction	Observations	Causal variables	What factors in the image account for the change in position?
Causal discovery	Causal variables	Causal graph	Does mass determine the change in position of an object?
Causal mechanisms learning	Causal graph	Causal mechanisms (structural equations)	How mass determine the change in position of an object?

traditional model-based methods and allow RL agents to adapt quickly to unseen environments or tasks. However, we may not have perfect knowledge of the causal variables a priori in reality. Sometimes, we must deal with high-dimensional and unstructured data like visual information. In this case, RL agents need to be able to extract causal representations from raw data (Schölkopf et al., 2021). Depending on the task, causal representations can be abstract concepts such as emotions and preferences, or they can be more concrete things such as physical objects.

The complete process of learning a causal model from data is known as causal learning (Peters et al., 2017). It is different from causal reasoning (Imbens & Rubin, 2015; Glymour et al., 2016), which only focuses on estimating specific causal effects given the causal model. Causal learning involves extracting causal features, discovering causality, and learning causal mechanisms. Table 5 briefly summarizes their characteristics. All three of these components are significant and deserve further investigation. A great deal of research has been done on causal discovery (Spirtes et al., 2000; Pearl, 2009b; Peters et al., 2017; Vowels et al., 2022), a process of recovering the causal structure of a set of variables from data, particularly concerning conditional independence tests (Spirtes et al., 2000; Sun et al., 2007; Hoyer et al., 2008; Zhang et al., 2011). Under certain assumptions, such as faithfulness, algorithms can identify the Markov equivalence class of the underlying causal graph from observational data. Combining causal discovery with RL allows an agent to actively gather interventional data from the environment in an interactive way. Therefore, an interesting line of research in this field focuses on how to use interventional data or a combination of observational and interventional data for efficient causal discovery (Addanki et al., 2020; Jaber et al., 2020; Brouillard et al., 2020; Zhu et al., 2022a).

As for causal feature extraction, also known as causal representation learning (Schölkopf et al., 2021; Wang & Jordan, 2022; Shen et al., 2022), one possible method is to use an autoencoder to learn latent variables from high-dimensional observations (Yang et al., 2021a; Eghbal-zadeh et al., 2021; Tran et al., 2022). These methods can approximatively recover causal representations and structures by virtue of carefully designed constraint terms. The whole process is analogous to embedding an SCM into the learner, implicitly binding causal discovery and causal mechanisms learning in one solution. When multiple environments or tasks are available, one may also obtain causal representations through mining invariance (Zhang et al., 2020a; Bica et al., 2021b; Saengkyongam et al., 2022) or clustering trajectories (Sontakke et al., 2021). However, it is still challenging to determine the number and granularity of causal variables required for policy learning, and the optimal causal representations often depend on the specific context and task. Overall, the development of causal learning methods suitable for decision-making problems is an underexplored area and has the potential to advance the RL community. Conversely, RL techniques may also contribute to the field of causal learning (Zhu et al., 2022a).

5.2 Causality-aware Multitask and Meta Reinforcement Learning

Multitask RL (Parisotto et al., 2015; Teh et al., 2017; D'Eramo et al., 2020; Vithayathil Varghese & Mahmoud, 2020) focuses on solving multiple decision tasks simultaneously, which is often encountered in robot manipulation. For instance, a robot may need to acquire various skills or achieve multiple goals. Meta-learning (Duan et al., 2016; Finn et al., 2017; Gupta et al., 2018; Xu et al., 2018), on the other hand, involves training on a task distribution to gain the ability to adapt quickly to a new task. Both of these approaches are crucial in the practical application of RL. Researchers have achieved impressive results without considering causality. A natural question arises: Is it still necessary to consider causality if we can train a high-capacity model with a diverse range of tasks? Different tasks are inherently different interventions in the data generation process (Schölkopf et al., 2021). Therefore, if we can involve sufficiently diverse interventions during the training phase, the resulting model may emerge with strong generalization ability to solve unseen tasks efficiently.

Recent research has provided evidence that large, pre-trained models perform well in tasks requiring few-shot learning or even zero-shot generalization ability (Brown et al., 2020; Wei et al., 2021). Thus, it is reasonable to assume that large decision models (Reed et al., 2022; Wen et al., 2022) may also possess this ability. Interestingly, Dasgupta et al. (2018) further demonstrated that the capability of causal inference might emerge from large-scale meta-RL. However, testing all potential interventions and their combinations in real-world situations is impractical. This is where causal modeling comes in. Causal modeling allows for the explicit incorporation of prior knowledge, enabling the model to align its understanding of the task or environment with human cognition. In addition, causal modeling divides the agent's knowledge into multiple independent and autonomous modules. A non-causal agent would have to re-learn all the modules, even for a slightly changed task, while a modular agent would only need to adopt a few modules to succeed in a new task, exhibiting a stronger knowledge transfer ability. This may also contribute to lifelong (or continual) learning (Xie & Finn, 2022; Khetarpal et al., 2022), allowing for fast adaptation to new tasks that arise in sequence.

5.3 Human-in-the-loop Learning and Reinforcement Learning from Human Feedback

Human-in-the-loop learning (HiLL) (Mosqueira-Rey et al., 2022) is a form of machine learning in which humans actively participate in the development cycle of machine learning models or algorithms. This can involve providing labels, preferences, or other types of feedback. When the data or task being learned is complex or requires high levels of cognition, HiLL often produces better results because humans can provide valuable insights or knowledge to the model that it may be difficult for the model to learn on its own (Zhang & Bareinboim, 2022).

In the context of RL, HiLL refers to involving humans in the MDP to replace the reward function that provides feedback signals. This allows us to train RL agents with the help of human knowledge and values rather than struggling to define a sophisticated reward function (Zhang & Bareinboim, 2022). This idea is closely related to RLHF (Reinforcement Learning from Human Feedback) (Christiano et al., 2017), a concept that has gained increasing attention recently in the training of large language models (Ziegler et al., 2020; Glaese et al., 2022; Ouyang et al., 2022), where human instructors provide rewards (or penalties) to a model to encourage (or discourage) certain behaviors. From a causal perspective, humans can provide machine learning models with a strong understanding of causality based on their knowledge of the world, which can help filter out behaviors that may lead to negative outcomes. However, it is important to note that humans and machines may have different observations or perceptions of the world, and non-causal-aware RL agents may be influenced by confounding variables (Gasse et al., 2021). In addition, we often need to consider the issue of limited budgets, as our goal is to provide meaningful feedback to RL agents at the lowest possible cost. Finally, in addition to scalar feedback, we may also provide more informative feedback to agents in the form of counterfactuals (Karalus & Lindner, 2022).

5.4 Theoretical Advances in Causal Reinforcement Learning

The vast majority of research in the field has focused on the MAB problem. Lattimore et al. (2016) first introduced the causal bandits problem, where the agent can observe variables other than the reward after performing an action. The agent is expected to use this information to infer the causal structure and identify the arm with the highest reward more efficiently. Sen et al. (2017a) later built upon this work by developing a gap-dependent regret bound. Yabe et al. (2018) expanded upon this setting by allowing the agent to intervene on multiple nodes. Lee & Bareinboim (2018) formally demonstrated that non-causal agents are suboptimal and investigated the impact of non-manipulable variables in subsequent work (Lee & Bareinboim, 2019). Lu et al. (2021) studied the problem of causal bandits with prior causal knowledge, while Nair et al. (2021) focused on the problem with budget constraints in different causal structures. More recently, Kroon et al. (2022) eliminated the need for a priori causal knowledge by introducing the concept of "separating sets" in causal discovery. Similarly, Bilodeau et al. (2022) derived the optimal regret when observing a d-separator. Additionally, some research has focused on controlling confounding variables in the bandits problem Bareinboim et al. (2015); Sen et al. (2017b).

The multi-step decision-making problem is more complex than the MAB problem as it involves state transitions. Some studies focused on the dynamic treatment regimes (Zhang & Bareinboim, 2019; Zhang, 2020), which can be modeled as an MDP with a global confounder variable. There have also been studies on confounded MDPs (Zhang & Bareinboim; Wang et al., 2021b), a more general problem that assumes the existence of unobserved confounders at each time step. Overall, RL empowered by causality can achieve better theoretical bounds than non-causal approaches. The causal bandits problem and its variants have received much attention within the research community, but the MDP problem has not been studied as extensively. Moreover, in addition to understanding how causal modeling enhances the regret bounds, some work studied the identifiability of causal effects (Zhang et al., 2020b; Lu et al., 2022) or the underlying causal structure (Huang et al., 2022a). How to bring together theoretical results and practical solutions to create more effective causal RL algorithms is an open issue that is worth further attention and investigation.

5.5 Benchmarking Causal Reinforcement Learning

In RL, we are typically interested in the efficiency and convergence of the algorithm. Atari 2600 Games and Mujoco locomotion tasks (Brockman et al., 2016) are commonly used as benchmarks for discrete and continuous control problems. There are also experimental environments that evaluate the generalizability and robustness of RL, such as Procgen (Cobbe et al., 2020). Some benchmarks focus on multitask learning, meta-learning, and curriculum learning for reinforcement learning, such as RLBench (James et al., 2019), Meta-World (Yu et al., 2021), Alchemy (Wang et al., 2021a), and Causal-World Ahmed et al. (2022). Among them, Causal-World delivers a diverse set of robotic manipulation tasks with a shared attribute set and structure that require the robot to construct a goal shape using the provided building blocks. This benchmark provides interfaces to manually modify objects' attributes (e.g., size, mass, and color), so researchers can intervene concerning these attributes in order to generate a series of tasks with the same causal structure.

Since causal RL is not limited to a particular type of problem, evaluation metrics may vary depending on the specific mission. While existing experimental environments have provided good benchmarks for evaluating algorithms with various metrics, the data generation processes within these environments are often opaque, hidden within game simulators or physics engines. This lack of transparency makes it difficult for researchers to fully understand the causal mechanisms behind the problems they are attempting to solve, hindering the development of the field. Recently, Ke et al. proposed a new set of environments focusing on causal discovery in visual-based RL, allowing researchers to specify the causal graph and adjust its complexity. However, as we demonstrated in section 4, much of the existing research still relies on toy environments to test the effectiveness of algorithms. Developing a suitable benchmark for causal RL remains an open question. In addition to the previously mentioned properties, a good benchmark should consider the multiple factors comprehensively, as discussed in section 4.4.

5.6 Real-world Causal Reinforcement Learning

Additionally, we note that there are currently very few real-world applications of causal RL. To make it more widely applicable, we must carefully address the various challenges posed by reality. Dulac-Arnold et al. (2020; 2021) identified nine critical challenges that are holding back the use of RL in the real world: limited samples; unknown and significant delays; high-dimensional input; safety constraints; partial observability; multi-objective or unspecified reward functions; low latencies; offline learning; and the need for explainability.

We have discussed some of these issues in paper, many of which are related to ignorance of causality. For instance, the challenge of learning from limited samples corresponds to the sample efficiency issue discussed in section 4.1. Learning from high-dimensional inputs and multiple reward functions relates to the generalization problem outlined in section 4.2. Offline learning raises concerns about spurious correlations (section 4.3), and security and explainability are covered in section 4.4.

Although causal modeling offers promising solutions to these real-world challenges, current experimental environments often fall short in meeting the research needs. As discussed in section 5.5, popular benchmarks are often treated as black boxes, and researchers have limited access to and understanding of the causal mechanisms by which these black boxes generate data. This situation significantly hinders the development of this research area. Therefore, establishing benchmarks that fully account for real-world factors would be valuable work that would also help increase the practical use of causal RL in the real world.

6 Conclusion

In summary, causal RL is a promising method of solving complex decision-making problems under uncertainty. It is an understudied but significant research direction. By explicitly modeling the causal structure of the target problem, causal RL algorithms can learn optimal policies more efficiently and make more informed decisions. In this survey, we aimed to clarify the terminologies and concepts related to causal RL and to establish connections between existing work. We proposed a problem-oriented taxonomy and systematically discussed and analyzed the latest advances in causal RL, focusing on how they address the four critical challenges facing RL.

While there is still much work to be done in this field, the results to date are encouraging. They suggest that causal RL has the potential to significantly improve the performance of RL systems in a wide range of problems. Here, we summarize the key conclusions of this survey.

- Causal RL is an emerging branch of RL that emphasizes understanding and utilizing causality to make better decisions.
- Causal modeling can enhance sample efficiency (section 4.1) and generalization ability (section 4.2); however, the causal variables may be unknown. In real-world scenarios, RL agents may need to learn about causal representation and environmental dynamics from raw data (section 5.1).
- In proper conditions, causal relationships are identifiable (section 5.4), i.e., RL agents can recover causal relationships from observed data. Additionally, multitask learning and meta-learning can facilitate causal learning (section 5.2); in turn, causality can improve the knowledge transfer ability to solve various tasks more effectively.
- Correlation does not imply causation. Spurious correlations can lead to a distorted understanding of the environment and task, resulting in suboptimal policies (section 4.3). Apart from relying on causal reasoning techniques, we can also utilize human understanding of causality to aid RL (section 5.3).
- In real-world applications, performance is not the only concern. Other factors, such as explainability, fairness, and security, must also be considered (section 4.4). Current benchmarks require greater transparency and a comprehensive, multi-faceted evaluation protocol for reinforcement learning (section 5.5), which has significant implications for advancing real-world applications of causal reinforcement learning (section 5.6).

We hope this survey will help establish connections between existing work in causal reinforcement learning, inspire further exploration and development, and provide a common ground and comprehensive resource for those looking to learn more about this exciting field.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pp. 22–31. PMLR, 2017.

Raghavendra Addanki, Shiva Kasiviswanathan, Andrew Mcgregor, and Cameron Musco. Efficient Intervention Design for Causal Discovery with Latents. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 63–73. PMLR, November 2020.

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wuthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning. In *International Conference on Learning Representations*, February 2022.

Eitan Altman. *Constrained Markov Decision Processes.* PhD thesis, INRIA, 1995.

Eitan Altman. *Constrained Markov Decision Processes: Stochastic Modeling.* Routledge, 1999.

Chenjia Bai, Lingxiao Wang, Yixin Wang, Zhaoran Wang, Rui Zhao, Chenyao Bai, and Peng Liu. Addressing Hindsight Bias in Multigoal Reinforcement Learning. *IEEE Transactions on Cybernetics*, pp. 1–14, 2021. ISSN 2168-2275. doi: 10.1109/TCYB.2021.3107202.

Sreejith Balakrishnan, Jianxin Bi, and Harold Soh. SCALES: From Fairness Principles to Constrained Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 46–55, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534190.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with Unobserved Confounders: A Causal Approach. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Andrew G. Barto. Intrinsic Motivation and Reinforcement Learning. In Gianluca Baldassarre and Marco Mirolli (eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems*, pp. 17–47. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-32375-1. doi: 10.1007/978-3-642-32375-1_2.

Jared J. Beard and Ali Baheri. Safety Verification of Autonomous Systems: A Multi-Fidelity Reinforcement Learning Approach. *arXiv preprint arXiv:2203.03451*, 2022.

Andrew Bennett, Nathan Kallus, Lihong Li, and Ali Mousavi. Off-policy Evaluation in Infinite-Horizon Reinforcement Learning with Latent Confounders. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 1999–2007. PMLR, March 2021.

Ioana Bica, Daniel Jarrett, Alihan Hüyük, and Mihaela van der Schaar. Learning "What-if" Explanations for Sequential Decision-Making. In *International Conference on Learning Representations*, March 2021a.

Ioana Bica, Daniel Jarrett, and Mihaela van der Schaar. Invariant Causal Imitation Learning for Generalizable Policies. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3952–3964. Curran Associates, Inc., 2021b.

Blair Bilodeau, Linbo Wang, and Daniel M. Roy. Adaptively Exploiting d-Separators with Causal Bandits. In *Advances in Neural Information Processing Systems*, October 2022.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym, June 2016.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable Causal Discovery from Interventional Data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21865–21877. Curran Associates, Inc., 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Lars Buesing, Theophane Weber, Yori Zwols, Nicolas Heess, Sebastien Racaniere, Arthur Guez, and Jean-Baptiste Lespiau. Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search. In *International Conference on Learning Representations*, February 2019.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, February 2022.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better Exploration with Optimistic Actor Critic. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging Procedural Generation to Benchmark Reinforcement Learning, July 2020.

Youri Coppens, Kyriakos Efthymiadis, Tom Lenaerts, Ann Nowé, Tim Miller, Rosina Weber, and Daniele Magazzeni. Distilling deep reinforcement learning policies in soft decision trees. In *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, pp. 1–6, 2019.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*, pp. 2185–2195. PMLR, 2020.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020. doi: 10.1145/3351095.3372878.

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal Reasoning from Meta-reinforcement learning. December 2018.

Zhihong Deng, Zuyue Fu, Lingxiao Wang, Zhuoran Yang, Chenjia Bai, Zhaoran Wang, and Jing Jiang. SCORE: Spurious COrrelation REduction for Offline Reinforcement Learning, October 2021.

Carlo D'Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, and Jan Peters. Sharing Knowledge in Multi-Task Deep Reinforcement Learning. In *International Conference on Learning Representations*, March 2020.

Wenhao Ding, Haohong Lin, Bo Li, and Ding Zhao. Generalizing Goal-Conditioned Reinforcement Learning with Variational Causal Reasoning. In *Advances in Neural Information Processing Systems*, October 2022.

Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning? In *International Conference on Learning Representations*, March 2020.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL\$$\hat{2}$\$: Fast Reinforcement Learning via Slow Reinforcement Learning. November 2016. doi: 10.48550/arXiv.1611.02779.

Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. An empirical investigation of the challenges of real-world reinforcement learning. *arXiv preprint arXiv:2003.11881*, 2020.

Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021. doi: 10.1007/s10994-021-05961-4.

Hamid Eghbal-zadeh, Florian Henkel, and Gerhard Widmer. LEARNING TO INFER UNSEEN CONTEXTS IN CAUSAL CONTEXTUAL REINFORCEMENT LEARNING. pp. 11, 2021.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(27):6435–6467, November 2021. ISSN 1573-0964. doi: 10.1007/s11229-021-03141-4.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, October 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00511.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1126–1135. PMLR, July 2017.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pp. 2974–2982, New Orleans, Louisiana, USA, February 2018. AAAI Press. ISBN 978-1-57735-800-8.

Andrew Forney, Judea Pearl, and Elias Bareinboim. Counterfactual Data-Fusion for Online Reinforcement Learners. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1156–1164. PMLR, July 2017.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12107.

Pratik Gajane, Akrati Saxena, Maryam Tavakol, George Fletcher, and Mykola Pechenizkiy. Survey on Fair Reinforcement Learning: Theory and Practice, May 2022.

Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. Causal Inference in Recommender Systems: A Survey and Future Directions. August 2022. doi: 10.48550/arXiv.2208.12397.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Maxime Gasse, Damien Grasset, Guillaume Gaudron, and Pierre-Yves Oudeyer. Causal Reinforcement Learning using Observational and Interventional Data, June 2021.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, September 2022.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019. doi: 10.3389/fgene.2019.00524.

Madelyn Glymour, Judea Pearl, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer.* John Wiley & Sons, 2016.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain Adaptation with Conditional Transferable Components. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 2839–2848. PMLR, June 2016.

Robert Grande, Thomas Walsh, and Jonathan How. Sample Efficient Reinforcement Learning with Gaussian Processes. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1332–1340. PMLR, June 2014.

Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *International Conference on Machine Learning*, pp. 1792–1801. PMLR, 2018.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A Review of Safe Reinforcement Learning: Methods, Theory and Applications. *arXiv preprint arXiv:2205.10330*, 2022.

Jiaxian Guo, Mingming Gong, and Dacheng Tao. A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning. In *International Conference on Learning Representations*, March 2022.

Abhishek Gupta, Russell Mendonca, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Meta-Reinforcement Learning of Structured Exploration Strategies. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Pouya Hamadanian, Malte Schwarzkopf, Siddartha Sen, and Mohammad Alizadeh. How Reinforcement Learning Systems Fail and What to do About It. In *The 2nd Workshop on Machine Learning and Systems (EuroMLSys)*, 2022.

Paul L. Harris, Tim German, and Patrick Mills. Children's use of counterfactual thinking in causal reasoning. *Cognition*, 61(3):233–259, December 1996. ISSN 0010-0277. doi: 10.1016/S0010-0277(96)00715-9.

Patrick Hart and Alois Knoll. Counterfactual Policy Evaluation for Decision-Making in Autonomous Driving, November 2020.

Tue Herlau and Rasmus Larsen. Reinforcement Learning of Causal Variables Using Mediation Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6910–6917, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i6.20648.

Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021. doi: 10.1016/j.knosys.2020.106685.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, pp. 1480–1490. PMLR, 2017.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. AdaRL: What, Where, and How to Adapt in Transfer Reinforcement Learning. In *International Conference on Learning Representations*, May 2022a.

Biwei Huang, Chaochao Lu, Liu Leqi, Jose Miguel Hernandez-Lobato, Clark Glymour, Bernhard Schölkopf, and Kun Zhang. Action-Sufficient State Representation Learning for Control with Structural Constraints. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 9260–9279. PMLR, June 2022b.

Wen Huang, Lu Zhang, and Xintao Wu. Achieving Counterfactual Fairness for Causal Bandit. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6952–6959, June 2022c. ISSN 2374-3468. doi: 10.1609/aaai.v36i6.20653.

Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press, 2015.

Kayoko Inagaki and Giyoo Hatano. Young Children's Understanding of the Mind-Body Distinction. *Child Development*, 64(5):1534–1549, 1993. ISSN 1467-8624. doi: 10.1111/j.1467-8624.1993.tb02969.x.

Alex Irpan. Deep Reinforcement Learning Doesn't Work Yet, 2018.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9551–9561. Curran Associates, Inc., 2020.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The Robot Learning Benchmark & Learning Environment, September 2019.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. Causal Direction of Data Collection Matters: Implications of Causal and Anticausal Learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9499–9513, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.748.

Nicholas K. Jong and Peter Stone. State Abstraction Discovery from Irrelevant State Variables. In *IJCAI*, volume 8, pp. 752–757. Citeseer, 2005.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Sham Machandranath Kakade. *On the Sample Complexity of Reinforcement Learning.* PhD thesis, University College London, March 2003.

Ken Kansky, Tom Silver, David A. Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, and Dileep George. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In *International Conference on Machine Learning*, pp. 1809–1818. PMLR, 2017.

Jakob Karalus and Felix Lindner. Accelerating the Learning of TAMER with Counterfactual Explanations, July 2022.

Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Rezende, Yoshua Bengio, Christopher Pal, and Michael Mozer. Systematic Evaluation of Causal Discovery in Visual Model Based Reinforcement Learning. pp. 13.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards Continual Reinforcement Learning: A Review and Perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, December 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13673.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A Survey of Generalisation in Deep Reinforcement Learning, January 2022.

Barbara Koslowski and Amy Masnick. The Development of Causal Reasoning. In Usha Goswami (ed.), *Blackwell Handbook of Childhood Cognitive Development*, pp. 257–281. Blackwell Publishers Ltd, Malden, MA, USA, January 2002. ISBN 978-0-470-99665-2 978-0-631-21840-1. doi: 10.1002/9780470996652.ch12.

Arnoud De Kroon, Joris Mooij, and Danielle Belgrave. Causal Bandits without prior knowledge using separating sets. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pp. 407–427. PMLR, June 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential Causal Imitation Learning with Unobserved Confounders. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14669–14680. Curran Associates, Inc., 2021.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Finnian Lattimore, Tor Lattimore, and Mark D. Reid. Causal bandits: Learning good interventions via causal inference. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 1189–1197, Red Hook, NY, USA, December 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9.

Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. Network Randomization: A Simple Technique for Generalization in Deep Reinforcement Learning. In *International Conference on Learning Representations*, March 2020.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6131–6141. PMLR, July 2021a.

Sanghack Lee and Elias Bareinboim. Structural Causal Bandits: Where to Intervene? In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Sanghack Lee and Elias Bareinboim. Structural Causal Bandits with Non-Manipulable Variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4164–4172, July 2019. doi: 10.1609/aaai.v33i01.33014164.

Tabitha E. Lee, Jialiang Alan Zhao, Amrita S. Sawhney, Siddharth Girdhar, and Oliver Kroemer. Causal Reasoning in Simulation for Structure and Transfer Learning of Robot Manipulation Policies. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4776–4782, May 2021b. doi: 10.1109/ICRA48506.2021.9561439.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Luofeng Liao, Zuyue Fu, Zhuoran Yang, Yixin Wang, Mladen Kolar, and Zhaoran Wang. Instrumental Variable Value Iteration for Causal Offline Reinforcement Learning, July 2021.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. Balancing Between Accuracy and Fairness for Interactive Recommendation with Reinforcement Learning. In Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin

Pan (eds.), *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pp. 155–167, Cham, 2020. Springer International Publishing. ISBN 978-3-030-47426-3. doi: 10.1007/978-3-030-47426-3_13.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Leon Bottou. Discovering Causal Signals in Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 58–66, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.14.

Chaochao Lu and José Miguel Hernández Lobato. Deconfounding Reinforcement Learning in Observational Settings. December 2018.

Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. Sample-Efficient Reinforcement Learning via Counterfactual-Based Data Augmentation, December 2020.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal Bandits with Unknown Graph Structure. In *Advances in Neural Information Processing Systems*, October 2021.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Efficient Reinforcement Learning with Prior Causal Knowledge. In *First Conference on Causal Learning and Reasoning*, February 2022.

Fan-Ming Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. A Survey on Model-based Reinforcement Learning, June 2022.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable Reinforcement Learning through a Causal Lens. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2493–2500, April 2020. ISSN 2374-3468. doi: 10.1609/aaai.v34i03.5631.

Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S. Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, and Remi Munos. Counterfactual Credit Assignment in Model-Free Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7654–7664. PMLR, July 2021.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, August 2022. ISSN 1573-7462. doi: 10.1007/s10462-022-10246-w.

Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented agents. *Advances in Neural Information Processing Systems*, 32, 2019.

Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal Induction from Visual Observations for Goal Directed Tasks, October 2019.

Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and Non-Budgeted Causal Bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 2017–2025. PMLR, March 2021.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy Policy Evaluation For Sequential Decisions Under Unobserved Confounding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18819–18831. Curran Associates, Inc., 2020.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving Rubik's Cube with a Robot Hand, October 2019.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) Efficient Reinforcement Learning via Posterior Sampling. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, October 2022.

Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Actor-Mimic: Deep Multitask and Transfer Reinforcement Learning. November 2015. doi: 10.48550/arXiv.1511.06342.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven Exploration by Self-supervised Prediction. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2778–2787. PMLR, July 2017.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, December 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.669.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009a. doi: 10.1214/09-SS057.

Judea Pearl. *Causality*. Cambridge university press, 2009b.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic books, 2018.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3803–3810. IEEE, 2018. doi: 10.1109/ICRA.2018.8460528.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 978-0-262-03731-0 978-0-262-34429-6.

Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual Data Augmentation using Locally Factored Dynamics. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3976–3990. Curran Associates, Inc., 2020.

Silviu Pitis, Elliot Creager, Ajay Mandlekar, and Animesh Garg. MoCoDA: Model-based Counterfactual Data Augmentation. In *Advances in Neural Information Processing Systems*, October 2022.

Erika Puiutta and Eric Veith. Explainable reinforcement learning: A survey. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 77–95. Springer, 2020. doi: 10.1007/978-3-030-57321-8_5.

Maithra Raghu, Alex Irpan, Jacob Andreas, Bobby Kleinberg, Quoc Le, and Jon Kleinberg. Can deep reinforcement learning solve Erdos-Selfridge-Spencer games? In *International Conference on Machine Learning*, pp. 4238–4246. PMLR, 2018.

Roberta Raileanu and Rob Fergus. Decoupling value and policy for generalization in reinforcement learning. In *International Conference on Machine Learning*, pp. 8787–8798. PMLR, 2021.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A Generalist Agent. *Transactions on Machine Learning Research*, November 2022.

Danilo J. Rezende, Ivo Danihelka, George Papamakarios, Nan Rosemary Ke, Ray Jiang, Theophane Weber, Karol Gregor, Hamza Merzic, Fabio Viola, Jane Wang, Jovana Mitrovic, Frederic Besse, Ioannis Antonoglou, and Lars Buesing. Causally Correct Partial Models for Reinforcement Learning, February 2020.

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

Richard M. Ryan and Edward L. Deci. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, 25(1):54–67, January 2000. ISSN 0361-476X. doi: 10.1006/ceps.1999.1020.

Sorawit Saengkyongam, Nikolaj Thams, Jonas Peters, and Niklas Pfister. Invariant Policy Learning: A Causal Perspective, September 2022.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Carolyn A Schult and Henry M Wellman. Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *Cognition*, 62(3):291–324, March 1997. ISSN 0010-0277. doi: 10.1016/S0010-0277(96)00786-X.

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal Influence Detection for Improving Efficiency in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, October 2021.

Rajat Sen, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. Identifying Best Interventions through Online Importance Sampling. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3057–3066. PMLR, July 2017a.

Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alex Dimakis, and Sanjay Shakkottai. Contextual Bandits with Latent Confounders: An NMF Approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 518–527. PMLR, April 2017b.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly Supervised Disentangled Generative Causal Representation Learning. *Journal of Machine Learning Research*, 23 (241):1–55, 2022. ISSN 1533-7928.

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally Regularized Learning with Agnostic Data Selection Bias. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 411–419, October 2018. doi: 10.1145/3240508.3240577.

Thomas R. Shultz. Rules of Causal Attribution. *Monographs of the Society for Research in Child Development*, 47(1):1–51, 1982. ISSN 0037-976X. doi: 10.2307/1165893.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, October 2017. ISSN 1476-4687. doi: 10.1038/nature24270.

Steven Sloman. *Causal Models: How People Think about the World and Its Alternatives.* Oxford University Press, 2005.

David Sobel and Jessica Sommerville. The Importance of Discovery in Children's Causal Learning from Interventions. *Frontiers in Psychology*, 1, 2010. ISSN 1664-1078. doi: 10.3389/fpsyg.2010.00176.

Sumedh A. Sontakke, Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9848–9858. PMLR, July 2021.

Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search.* MIT Press, 2000. ISBN 978-0-262-19440-2.

James H. Stock and Francesco Trebbi. Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194, 2003. doi: 10.1257/089533003769204416.

Jianyu Su, Stephen Adams, and Peter A. Beling. Counterfactual Multi-Agent Reinforcement Learning with Graph Convolution Communication, December 2020.

Xiaohai Sun, Dominik Janzing, Bernhard Schölkopf, and Kenji Fukumizu. A kernel-based causal learning algorithm. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 855–862, New York, NY, USA, June 2007. Association for Computing Machinery. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273604.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* MIT press, 2018.

Gokul Swamy, Sanjiban Choudhury, Drew Bagnell, and Steven Wu. Causal Imitation Learning under Temporally Correlated Noise. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 20877–20890. PMLR, June 2022.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased Scene Graph Generation From Biased Training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3713–3722, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00377.

Yee Teh, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30. IEEE, 2017. doi: 10.1109/IROS.2017.8202133.

Thien Q. Tran, Kazuto Fukuchi, Youhei Akimoto, and Jun Sakuma. Unsupervised Causal Binary Concepts Discovery with VAE for Black-Box Model Explanation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9614–9622, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i9.21195.

Stelios Triantafyllou, Adish Singla, and Goran Radanovic. 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*, AIES '22, pp. 739–752, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534133.

Stratis Tsirtsis, Abir De, and Manuel Rodriguez. Counterfactual Explanations in Sequential Decision Making Under Uncertainty. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30127–30139. Curran Associates, Inc., 2021.

Nelson Vithayathil Varghese and Qusay H. Mahmoud. A survey of multi-task deep reinforcement learning. *Electronics*, 9(9):1363, 2020. doi: 10.3390/electronics9091363.

Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Computing Surveys*, 55(4):82:1–82:36, November 2022. ISSN 0360-0300. doi: 10.1145/3527154.

Jane X. Wang, Michael King, Nicolas Porcel, Zeb Kurth-Nelson, Tina Zhu, Charlie Deck, Peter Choy, Mary Cassin, Malcolm Reynolds, Francis Song, Gavin Buttimore, David P. Reichert, Neil Rabinowitz, Loic Matthey, Demis Hassabis, Alexander Lerchner, and Matthew Botvinick. Alchemy: A benchmark and analysis toolkit for meta-reinforcement learning agents, October 2021a.

Kaixin Wang, Bingyi Kang, Jie Shao, and Jiashi Feng. Improving generalization in reinforcement learning with mixture regularization. *Advances in Neural Information Processing Systems*, 33:7968–7978, 2020a.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably Efficient Causal Reinforcement Learning with Confounded Observational Data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21164–21175. Curran Associates, Inc., 2021b.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual Commonsense R-CNN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10757–10767, Seattle, WA, USA, June 2020b. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.01077.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal Attention for Unbiased Visual Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3091–3100, 2021c.

Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking Model-Based Reinforcement Learning, July 2019.

Yixin Wang and Michael I. Jordan. Desiderata for Representation Learning: A Causal Perspective, February 2022.

Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal Dynamics Learning for Task-Independent State Abstraction. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 23151–23180. PMLR, June 2022.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models Are Zero-Shot Learners. September 2021. doi: 10.48550/arXiv.2109.01652.

Henry M. Wellman. *The Child's Theory of Mind.* The Child's Theory of Mind. The MIT Press, Cambridge, MA, US, 1992. ISBN 978-0-262-23153-4 978-0-262-73099-0.

Zac Wellmer and James T. Kwok. Dropout's dream land: Generalization from learned simulators to reality. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 255–270. Springer, 2021.

Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1152. PMLR, 2021.

Ying Wen, Ziyu Wan, Ming Zhou, Shufang Hou, Zhe Cao, Chenyang Le, Jingxiao Chen, Zheng Tian, Weinan Zhang, and Jun Wang. On Realization of Intelligent Decision-Making in the Real World: A Foundation Decision Model Perspective, December 2022.

Xiaobao Wu, Chunping Li, and Yishu Miao. Discovering Topics in Long-tailed Corpora with Causal Intervention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 175–185, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.15.

Annie Xie and Chelsea Finn. Lifelong Robotic Reinforcement Learning by Retaining Experiences. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, pp. 838–855. PMLR, November 2022.

Zhongwen Xu, Hado P van Hasselt, and David Silver. Meta-Gradient Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. Causal Bandits with Propagating Inference. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5512–5520. PMLR, July 2018.

Chao-Han Huck Yang, I.-Te Danny Hung, Yi Ouyang, and Pin-Yu Chen. Training a Resilient Q-network against Observational Interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8814–8822, June 2022a. doi: 10.1609/aaai.v36i8.20862.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9588–9597, June 2021a. doi: 10.1109/CVPR46437.2021.00947.

Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Jianye Hao, Zhaopeng Meng, Peng Liu, and Zhen Wang. Exploration in Deep Reinforcement Learning: A Comprehensive Survey, July 2022b.

Yijun Yang, Jing Jiang, Tianyi Zhou, Jie Ma, and Yuhui Shi. Pareto Policy Pool for Model-based Offline Reinforcement Learning. In *International Conference on Learning Representations*, 2021b.

Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. In *International Conference on Learning Representations*, 2021.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Avnish Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning, June 2021.

Yang Yu. Towards sample efficient reinforcement learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 5739–5743, Stockholm, Sweden, July 2018. AAAI Press. ISBN 978-0-9992411-2-7.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification, July 2015.

Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, Murray Shanahan, Victoria Langston, Razvan Pascanu, Matthew Botvinick, Oriol Vinyals, and Peter Battaglia. Deep reinforcement learning with relational inductive biases. In *International Conference on Learning Representations*, 2019.

Amy Zhang, Nicolas Ballas, and Joelle Pineau. A dissection of overfitting and generalization in continuous reinforcement learning. *arXiv preprint arXiv:1806.07937*, 2018a.

Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant Causal Prediction for Block MDPs. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11214–11224. PMLR, November 2020a.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *International Conference on Learning Representations*, February 2022.

Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018b.

Junzhe Zhang. Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11012–11022. PMLR, November 2020.

Junzhe Zhang and Elias Bareinboim. Markov Decision Processes with Unobserved Confounders: A Causal Approach. pp. 40.

Junzhe Zhang and Elias Bareinboim. Transfer Learning in Multi-Armed Bandits: A Causal Approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 1340–1346, Melbourne, Australia, August 2017. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/186.

Junzhe Zhang and Elias Bareinboim. Fairness in Decision-Making — The Causal Explanation Formula. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. doi: 10.1609/aaai.v32i1.11564.

Junzhe Zhang and Elias Bareinboim. Near-Optimal Reinforcement Learning in Dynamic Treatment Regimes. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Junzhe Zhang and Elias Bareinboim. Can Humans Be out of the Loop? In *Proceedings of the First Conference on Causal Learning and Reasoning*, pp. 1010–1025. PMLR, June 2022.

Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pp. 12263–12274, Red Hook, NY, USA, December 2020b. Curran Associates Inc. ISBN 978-1-71382-954-6.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pp. 804–813, Arlington, Virginia, USA, July 2011. AUAI Press. ISBN 978-0-9749039-7-2.

Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. doi: 10.1609/aaai.v29i1.9542.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep Stable Learning for Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5372–5382, 2021a.

Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 11–20, New York, NY, USA, July 2021b. Association for Computing Machinery. ISBN 978-1-4503-8037-9. doi: 10.1145/3404835.3462875.

Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 2980–2991, New York, NY, USA, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8312-7. doi: 10.1145/3442381.3449788.

Hanhan Zhou, Tian Lan, and Vaneet Aggarwal. PAC: Assisted Value Factorization with Counterfactual Predictions in Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, October 2022.

Deyao Zhu, Li Erran Li, and Mohamed Elhoseiny. CausalDyna: Improving Generalization of Dyna-style Reinforcement Learning via Counterfactual-Based Data Augmentation. November 2021.

Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal Discovery with Reinforcement Learning. In *International Conference on Learning Representations*, February 2022a.

Zheng-Mao Zhu, Xiong-Hui Chen, Hong-Long Tian, Kun Zhang, and Yang Yu. Offline Reinforcement Learning with Causal Structured World Models, June 2022b.

Zhuangdi Zhu, Kaixiang Lin, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *arXiv preprint arXiv:2009.07888*, 2020.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences, January 2020.