A Survey on Causal Reinforcement Learning

Yan Zeng, Ruichu Cai, *Senior Member, IEEE*, Fuchun Sun, *Fellow, IEEE*, Libo Huang, Zhifeng Hao, *Senior Member, IEEE*

Abstract—While Reinforcement Learning (RL) achieves tremendous success in sequential decision-making problems of many domains, it still faces key challenges of data inefficiency and the lack of interpretability. Interestingly, many researchers have leveraged insights from the causality literature recently, bringing forth flourishing works to unify the merits of causality and address well the challenges from RL. As such, it is of great necessity and significance to collate these Causal Reinforcement Learning (CRL) works, offer a review of CRL methods, and investigate the potential functionality from causality toward RL. In particular, we divide existing CRL approaches into two categories according to whether their causality-based information is given in advance or not. We further analyze each category in terms of the formalization of different models, ranging from the Markov Decision Process (MDP), Partially Observed Markov Decision Process (POMDP), Multi-Arm Bandits (MAB), and Dynamic Treatment Regime (DTR). Moreover, we summarize the evaluation matrices and open sources, while we discuss emerging applications, along with promising prospects for the future development of CRL.

Index Terms—Causal reinforcement learning, causal discovery, causal inference, Markov decision process, sequential decision making.

I. INTRODUCTION

REINFORCEMENT Learning (RL) is a general framework for an agent to learn a policy (the mapping function from states to actions) which maximizes the expected rewards in an environment [1]–[3]. It attempts to tackle sequential decision-making problems through the scheme of trial and error while the agent interacts with the environment. Due to its remarkable success in performance, it has been rapidly developed and deployed in various real-world applications, including games [4]–[6], robotics controlling [7], [8], and recommender systems [9], [10], etc., gaining increasing attention from researchers of different disciplines.

However, there are some key challenges of reinforcement learning that still need to be addressed. For instance, (i) *data inefficiency*. Previous methods are mostly hungry for interaction data, whereas in real-world scenarios, e.g., in medicine or healthcare [11], only a few recorded data are

Yan Zeng is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: yanazeng013@tsinghua.edu.cn).

Ruichu Cai is with the School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: cairuichu@gmail.com).

Fuchun Sun is with the Department of Computer Science and Technology, Beijing National Research Centre for Information Science and Technology, Tsinghua University, Beijing 100084, China (Corresponding author, e-mail: fcsun@tsinghua.edu.cn).

Libo Huang is with Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China (e-mail: www.huanglibo@gmail.com).

Zhifeng Hao is with the College of Science, Shantou University, Shantou 515063, China (e-mail: haozhifeng@stu.edu.cn).

available, majorly due to the costly, unethical or difficult collection procedures. (ii) *Lack of interpretability*. Existing methods often formalize the reinforcement learning problem through deep neural networks, which belong to the black boxes, taking sequential data as inputs and the policy as output. It is hard for them to uncover the internal relationships between states, actions, or rewards behind the data, and to offer intuitions about the policy characteristics. Such a challenge would impede the actual applications in industrial engineering.

Interestingly, causality may play an indispensable role in handling those aforementioned challenges from reinforcement learning [12], [13]. Causality considers two fundamental questions [14]: (1) What empirical evidence is required for legitimate inference of cause-effect relationships? This procedure of uncovering cause-effect relationships with evidence is briefly termed causal discovery. (2) Given the accepted causal information about a phenomenon, what inferences can we draw from such information, and how? Such a procedure of inferring causal effects or other interests is termed causal inference. Causality could empower the agents to do interventions or reason counterfactually via the *ladder of causation*, relaxing the requirements of a large amount of training data; it also enables to characterize the world model, potentially giving interpretability to how the agents interact with the environment.

In the past decades, both causality and reinforcement learning have achieved tremendous theoretical and technical development independently, whereas they could have been reconcilably integrated with each other [15]. Bareinboim [15] developed a unified framework called causal reinforcement learning by putting them under the same conceptual and theoretical umbrella and provided a tutorial for introduction online; Lu [16] was motivated by the current development in healthcare and medicine, and thus integrated causality and reinforcement learning, introducing causal reinforcement learning as well as emphasizing its potential applicability. Recently, a series of research related to causal reinforcement learning has been proposed, which needs a comprehensive survey on its development and applications. Thus, in this paper, we focus on providing our readers with good knowledge about concepts, categories, and practical issues of causal reinforcement learning.

Though there existed some relative reviews, i.e., Grimbly et al., [17] surveyed on causal multi-agent reinforcement learning; Bannon et al., [18] on causal effects' estimation and off-policy evaluation in batch reinforcement learning, here we consider cases but not limited to multi-agent or off-policy evaluation ones. Recently, Kaddour et al., [19] have uploaded a survey on causal machine learning on arXiv, including a chapter on causal reinforcement learning. They summarized the approaches according to different RL problems that causality

could bring benefits to, e.g., causal bandits, model-based RL, off-policy policy evaluation, etc. Such a taxonomic way may not be complete or integrated, which misses some other RL problems, e.g., multi-agent RL [18]. In this paper we only but completely construct a taxonomy framework for those causal reinforcement learning approaches. Our contributions of this survey paper are as follows:

- We formally define causal reinforcement learning, and to the best of our knowledge, we distinguish existing approaches into two categories from the perspective of causality for the first time. The first category is based on the prior causal information, where generally such methods assume the causal structure regarding to the environment or task is given a prior from the experts, while the second category is based on the unknown causal information, where the relative causal information has to be learned for the policy.
- We offer a comprehensive review of current methods on each category with systematical descriptions (and sketch maps). Regarding the first category, CRL methods take the best of prior causal information in policy learning, to improve sample efficiency, causal explanation or generalization capability. Regarding CRL with unknown causal information, these methods usually contain two phases: causal information learning and policy learning, which are carried out iteratively or in turn.
- We further analyze and discuss applications of CRL, evaluation metrics, open sources as well as future directions [1].

II. PRELIMINARY

Here we give basic definitions and concepts in both causality and RL, which will be utilized throughout the paper.

A. Causality

As for causality, we first define the notations and assumptions, and then introduce general methods from perspectives of causal discovery and causal inference. The former addresses the problem of identifying causal structures from purely observational data, which exploits statistical properties to tell the cause-effect relations between variables of interest; while the latter aims at inferring causal effects or other statistical interests when the causal relations are completely or partially known.

- *Definitions and assumptions*

*Definition 1 (**Structural Causal Model (SCMs) [14]**):* Given a set of equations in Eq.(1), if each equation represents an autonomous mechanism and each mechanism determines the value of only one distinct variable, then the model is called a structural causal model or causal model for short,

$$x_i = f_i(pa_i, u_i), i = 1, ..., n, \quad (1)$$

where x_i, u_i are the ith random variable and its error variable, respectively. f_i stands for the generation mechanism while pa_i is a set of x_i's parental variables, i.e., those immediate causes of x_i.

[1]Please see Appendix for evaluation metrics and open sources.

Given a SCM with two random variables, which satisfies,

$$x_1 = u_1, x_2 = f_2(x_1, u_2), \quad (2)$$

then $x_1 \rightarrow x_2$ is referred to as a causal graph [13]. Here, x_1 is the cause or parental parent of x_2. Every causal model can be associated with a directed graph.

*Definition 2 (**Rubin Causal Model [20], [21]**):* Rubin causal model involves an observational dataset of $\{Y_i, T_i, X_i\}$, where Y_i denotes a potential outcome of unit i; $T_i \in \{0, 1\}$ is an indicator variable for taking a treatment or not; and X_i is a set of covariates.

The Rubin causal model is also well-known as the potential outcome framework or the Neyman-Rubin potential outcomes. Since an individual unit cannot take different treatments simultaneously, but rather only one treatment at a time, it is impossible to obtain both potential outcomes and one has to estimate the missing one. With potential outcomes, the Rubin causal model aims at estimating the treatment effect.

*Definition 3 (**Treatment effect**):* Treatment effects can be evaluated at the population, treated group, subgroup, and individual levels. As for the population level, the Average Treatment Effect (ATE) is defined as

$$\text{ATE} = \mathbb{E}[Y(T = 1) - Y(T = 0)], \quad (3)$$

where $Y(T = 1)$ and $Y(T = 0)$ represent the potential outcomes with and without treatment, respectively. As for the treated group level, the Average Treatment effect on the Treated group (ATT) is defined as

$$\text{ATT} = \mathbb{E}[Y(T = 1) \mid T = 1] - \mathbb{E}[Y(T = 0) \mid T = 1], \quad (4)$$

where $Y(T = 1) \mid T = 1$ and $Y(T = 0) \mid T = 1$ represent the potential outcomes with and without treatment, respectively, in the treated group. As for the subgroup level, the Conditional Average Treatment Effect (CATE) is defined as

$$\text{CATE} = \mathbb{E}[Y(T = 1) \mid X = x] - \mathbb{E}[Y(T = 0) \mid X = x], \quad (5)$$

where $Y(T = 1) \mid X = x$ and $Y(T = 0) \mid X = x$ represent the potential outcomes with and without treatment, respectively, of the subgroup with $X = x$. As for the individual level, the Individual Treatment Effect (ITE) is defined as

$$\text{ITE}_i = Y_i(T = 1) - Y_i(T = 0), \quad (6)$$

where $Y_i(T = 1)$ and $Y_i(T = 0)$ represent the potential outcomes of the unit i with and without treatment, respectively.

*Definition 4 (**Confounder [22]**):* A confounder in a causal graph is the unobserved direct common cause of two observed variables. Especially, confounders in the potential outcome framework are those variables that affect both the treatment and the outcome.

*Definition 5 (**Instrumental Variables (IVs) [14]**):* An variable Z is an IV relative to the pair (T, Y), if it satisfies the following two conditions: i) Z is independent of all variables (including error terms) that have an influence on the outcome Y that is not mediated by T; ii) Z is not independence of the treatment T.

*Definition 6 (**Conditional Independence [14]**):* Let $\mathbf{X} = \{x_1, ..., x_n\}$ be a finite set of variables. Let $P(\cdot)$ be a joint

probability function over the variables in $\mathbf{X}$, and let $\mathbf{Y}, \mathbf{W}, \mathbf{Z}$ stand for any three subsets of variables in $\mathbf{X}$. The sets $\mathbf{Y}$ and $\mathbf{W}$ are said to be conditionally independent given $\mathbf{Z}$ if

$$P(y \mid w, z) = P(y \mid z), \text{whenever} P(w, z) > 0. \quad (7)$$

In words, learning the value of W does not provide additional information about Y, once we know Z, represented as $\mathbf{Y} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z}$.

One of the most widely applicable tools for conditional independence is Kernel-based Conditional Independence test (KCI-test), whose test statistic is calculated from the kernel matrices, characterizing the uncorrelatedness of functions in reproducing kernel Hilbert spaces [23].

*Definition 7 (**Back-Door** [14]):* A set of variables $\mathbf{Z}$ satisfies the back-door criterion relative to an ordered pair of variables (x_i, x_j) in a Directed Acyclic Graph (DAG) if: (i) no node in $\mathbf{Z}$ is a descendant of x_i; and (ii) $\mathbf{Z}$ blocks every path between x_i and x_j that contains an arrow into x_i. Similarly, if $\mathbf{Y}$ and $\mathbf{W}$ are two disjoint subsets of nodes in the DAG, then $\mathbf{Z}$ is said to satisfy the back-door criterion relative to $(\mathbf{Y}, \mathbf{W})$ if it satisfies the criterion relative to any pair (x_i, x_j) such that $x_i \in \mathbf{Y}$ and $x_j \in \mathbf{W}$.

*Definition 8 (**Front-Door** [14]):* A set of variables $\mathbf{Z}$ is said to satisfy the front-door criterion relative to an ordered pair of variables (x_i, x_j) if: (i) $\mathbf{Z}$ intercepts all directed paths from x_i to x_j; (ii) there is no back-door path from x_i to $\mathbf{Z}$; and (iii) all back-door paths from $\mathbf{Z}$ to x_j are blocked by x_i.
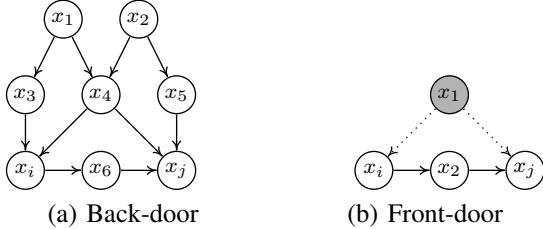

(a) Back-door   (b) Front-door

Fig. 1. Illustration examples for back-door and front-door criteria, where unshaded variables are observed while the shaded one is unobserved. x_1 in (b) is a latent confounder.

The back-door and front-door criteria are two simple graphical tests to tell whether a set of variables $\mathbf{Z} \subseteq \mathbf{X}$ is sufficient to estimate the causal effect $P(x_j \mid x_i)$. As illustrated in Fig.1, the set of variables $\mathbf{Z} = \{x_3, x_4\}$ satisfies the back-door criterion, while $\mathbf{Z} = \{x_2\}$ satisfies the front-door criterion.

The ladder of causality [24] has gained much attention from many researchers of various domains. It states there are three types of levels for causation, i.e., association, intervention and counterfactuals. They correspond to different interactions with the data generation process:

*Definition 9 (**Association** (seeing) [14]):* Let $\mathbf{Y}$ and $\mathbf{Z}$ be two subsets of variables in $\mathbf{X}$. The association sentence "what the value of $\mathbf{Z}$ is, if $\mathbf{Y}$ is found as y?" is interpreted as the causal effects of $\mathbf{Z}$ to $\mathbf{Y}$, denoted by $P(\mathbf{Z} \mid \mathbf{Y})$.

*Definition 10 (**Intervention** (doing) [14]):* Let $\mathbf{Y}$ and $\mathbf{Z}$ be two subsets of variables in $\mathbf{X}$. The intervention sentence "what the value of $\mathbf{Z}$ would obtain, if $\mathbf{Y}$ is assigned as y?" is interpreted as the causal effects of $\mathbf{Z}$ to an action $do(\mathbf{Y} = y)$, denoted by $P(\mathbf{Z} \mid do(\mathbf{Y} = y))$.

*Definition 11 (**Counterfactuals** (imagining) [14]):* Let $\mathbf{Y}$ and $\mathbf{Z}$ be two subsets of variables in $\mathbf{X}$. The counterfactual sentence "what the value of $\mathbf{Z}$ would have obtained, had $\mathbf{Y}$ been y?" is interpreted as the potential response of $\mathbf{Z}$ to an action $do(\mathbf{Y} = y)$, which is the solution of $\mathbf{Z}$ with the set of equations $f_{\mathbf{Z}}(\cdot)$, denoted by $P(\mathbf{Z}_{\mathbf{Y}} \mid y)$.

Assumptions 1-3 below are usually made for causal discovery to find the causal structure:

*Assumption 1 (**Causal Markov Assumption** [14]):* A necessary and sufficient condition for a probability population distribution P to be Markov relative to a causal graph (DAG) is that every variable is independent of all its nondescendants conditional on it parents.

*Assumption 2 (**Causal Faithfulness Assumption** [14]):* The probability distribution P in the population has no additional conditional independence relations that are not entailed by d-separation[2] to the causal graph.

*Assumption 3 (**Causal Sufficiency Assumption** [25]):* As for a set of variables $\mathbf{X}$, there is no hidden common cause, i.e., latent confounder, that causes more than one variable in $\mathbf{X}$.

Assumptions 4-6 are commonly derived for causal inference to estimate the treatment effect:

*Assumption 4: (**Stable Unit Treatment Value Assumption** [21]):* The potential outcomes of any given unit do not vary with the treatments assigned to other units, and for each unit, there are no different versions of treatment, which lead to different potential outcomes.

*Assumption 5 (**Ignorability** [21]):* Given the background covariates X, treatment assignment T is independent to the potential outcomes, i.e., $T \perp\!\!\!\perp Y(T = 0), Y(T = 1) \mid X$.

*Assumption 6 (**Positivity** [26]):* Given any value of X, the treatment assignment T is not deterministic:

$$P(T = t \mid X = x) > 0, \quad \forall \ t \text{ and } x. \quad (8)$$

- *Causal discovery*

As for identifying causal structures from the data, a traditional way is to use interventions, randomized or control experiments, which is in many situations too expensive, too time-consuming, or even too unethical to conduct [13]. Hence, uncovering causal information from purely observational data, known as causal discovery, has attracted much attention [22], [27], [28]. There are roughly two types of classic causal discovery approaches: constraint-based approaches and score-based ones. In the early of 1990's, constraint-based approaches exploited conditional independence relationships to recover the underlying causal structure among the observed variables, under appropriate assumptions. Such approaches include PC [29], and Fast Causal Inference (FCI) [14], [30], which allow different types of data distributions and causal relations, and could give asymptotically correct results. PC algorithm assumes that there is no latent confounder in the underlying causal graph; while FCI enables to handle the cases with latent confounders. However, what they recover belongs to an equivalence class of causal structures, which contains

---

[2]A set $\mathbf{W}$ is said to d-separate $\mathbf{Y}$ and $\mathbf{Z}$ if and only if $\mathbf{W}$ blocks every path from a variable in $\mathbf{Y}$ to a variable in $\mathbf{Z}$ [14].

multiple DAGs entailing the same conditional independence relationships. On the other hand, the score-based approaches attempt to search for the equivalence class via optimizing a properly defined score function [31], [32], such as the Bayesian Information Criterion (BIC) [33], the generalized score functions [27], etc. And they output one or multiple candidate causal graphs with the highest score. A well-known two-phase search procedure is Greedy Equivalence Search (GES) that searches directly over the space of equivalence classes [22], [31].

To distinguish different DAGs in the equivalence class and enjoy the unique identifiability of the causal structure, algorithms based on the constrained functional causal models have emerged [34]. These algorithms assumed the data generation mechanism, including the model class or noises' distribution: the effect variable is a function of direct causes along with an independent noise, as illustrated in Eq.(1) where causes pa_i are independent with the noise u_i. This gives rise to the causal structure's unique identifiability, in that the model assumptions, e.g., the independence between pa_i and u_i, only hold for the true causal direction but are violated for the wrong direction. Examples of these constrained functional causal models are linear non-Gaussian acyclic models (LiNGAM, [28]), additive noise models (ANM, [35]), post-nonlinear models (PNL, [36]), etc.

Furthermore, it was noted that there existed many significant but challenging topics that interest researchers. For instance, one may be interested in algorithms for the time series data. These algorithms include tsFCI [37], SVAR-FCI [38], tsLiNGAM [39], LPCMCI [40], etc. Especially, Granger causality allows to infer causal structure of time series, with no instantaneous effects or latent confounders[3]. It has been previously and commonly applied in prediction of economics [41]. Constraint-based causal Discovery from heterogeneous/NOnstationary Data (CD-NOD) works on the cases where the underlying generation process changes across domains or over time [34]. It uncovers the causal skeleton and directions, and estimates a low-dimensional representation for the changing causal modules.

- *Causal inference*

As for learning causal effects from the data, the most effective way is also to perform randomized experiments to compare the difference between the control and treatment groups. However, due to the high cost, practical and ethical issues, its applications are largely limited. Hence, estimating the treatment effect from observational data has gained increasing interests [20], [26].

Difficulties of causal inference from observational data lie in the presence of confounder variables, which leads to i) the selection bias between the treated and control groups, and ii) spurious effects. These problems would deteriorate the estimation performances of treatment outcomes. To handle the problem of spurious effects, a representative method is the stratification, also known as subclassification or blocking [42]. The idea is to split the entire group into homogeneous sub-

---

[3]Note that apart from tsLiNGAM and Granger causality, other mentioned methods from time series allow the presence of latent confounders.

---

groups, each of which in the treated and control groups shares similar characteristics over certain covariates [26]. To conquer the selection bias challenge, there are generally two types of causal inference approaches [26]. The first one aims at creating a pseudo group which is approximately consistent to the treated group. Such approaches include sample re-weighting methods [43], matching methods [44], tree-based methods [45], and representation based methods [46], etc. The other type of approaches such as meta-learning based ones, first train the outcome estimation model on the observed data and then correct the estimation bias stemmed from the selection bias [47], [48].

The above-mentioned causal inference methods rely on the satisfactions of Assumption 4-6. In practice, such assumptions may not always hold. For instance, when latent confounders exist, Assumption 5 does not hold, i.e., $T \not\perp\!\!\!\perp Y(T = 0), Y(T = 1) \mid X$. In such a case, one solution is to apply sensitivity analysis to investigate how inferences might change with various magnitudes of given unmeasured confounders. Sensitivity analysis quantifies the unmeasured confounding or hidden bias generally via differences between the unidentifiable distribution $P(Y(T = t) \mid T = 1 - t, X)$ and the identifiable distribution $P(Y(T = t) \mid T = t, X)$,

$$
\begin{aligned}
c_t(X) = {}& \mathbb{E}(Y(T = t) \mid T = 1 - t, X) \\
& - \mathbb{E}(Y(T = t) \mid T = t, X).
\end{aligned} \tag{9}
$$

Specifying bounds for $c_t(X)$, one could obtain the bounds for the expectation of outcomes $\mathbb{E}(Y(T = t))$, in the form of unidentifiable selection bias [49], [50]. Another possible solutions are to make the best of Instrumental Variables (IV) regression methods and Proximal Causal Learning (PCL) methods. These methods are used to predict the causal effects of the treatment or the policy, even with the existence of latent confounders [51]–[54]. It is worthwhile to note that the intuition behind PCL is to construct two conditionally independent proxy variables, so as to reflect the the effects of unobserved confounders [53], [54]. Figure 2 illustrates examples of the IV and proxy variables.
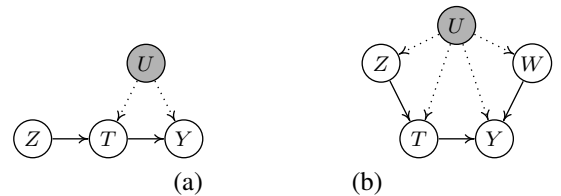

Fig. 2. Illustration examples for the instrumental variable and proxy variables, where unshaded variables are observed while the shaded one U is unobserved. T stands for the treatment, while Y is the outcome. In (a), Z is an instrumental variable while $\{Z, W\}$ are proxy variables in (b).

B. Reinforcement Learning

In this section, we first offer RL's characteristics and basic concepts, following the standard textbook definitions [55]. And thereafter we give baselines from perspectives of model-free and model-based RL methods, which implies that the world model (including dynamics and reward functions) is utilized or not.

● *Definitions*

Compared with supervised and unsupervised learning, RL enjoys advantages from two key components: optimal control, and trial and error. Based on the optimal control problem, Richard Bellman developed a dynamic programming approach, using the value function with the system's state information to give mathematical formalization [56], [57]. Such a value function is known as the Bellman equation, given as,

$$V(s_t) = r(s_t) + \gamma \sum P(s_{t+1} \mid s_t, a_t) \cdot V(s_{t+1}), \quad (10)$$

where $V(s_t)$ is the value function of the state s_t at time t, s_{t+1} is the next state, $r(s_t)$ is the reward function, and γ is a discount factor. $\mathcal{P}(s_{t+1}|s_t, a_t)$ is the transition probability of s_{t+1} given the current state s_t and action a_t. Learning through interactions is the essence of RL. An agent interacts with the environment by taking an action a_t in state s_t, and once observing its next state s_{t+1} and the reward $r(s_t)$, it needs to adjust the policy to strive for the optimal returns [1]–[3]. Such a trial-and-error learning regime is originated from animal psychology, which implies that actions accounting for good outcomes are likely to be repeated, which those for bad outcomes are muted [57], [58].

RL addresses the problem of learning a policy from the available information in different settings, including Multi-Armed Bandits (MAB), Contextual Bandits (CB), Markov Decision Process (MDP), Partially Observed Markov Decision Process (POMDP), Imitation Learning (IL), and Dynamic Treatment Regime (DTR).

*Definition 12 (**Markov Decision Process (MDP)** [55]):* A Markov decision process is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is a set of states $s \in \mathcal{S}$ and $\mathcal{A}$ is a set of actions $a \in \mathcal{A}$. $\mathcal{P}$ denotes the state transition probability that defines the distribution $\mathcal{P}(s_{t+1}|s_t, a_t)$, while $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the reward function. $\gamma \in [0, 1]$ is a discount factor.

*Definition 13: (**Partially Observed Markov decision Process (POMDP)** [55]):* The partially observed Markov decision process is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \mathcal{E}, \gamma)$, where $\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma$ are the same notations as MDP. $\mathcal{O}$ denotes the set of observations $o \in \mathcal{O}$ while $\mathcal{E}$ is an emission function that determines the distribution $\mathcal{E}(o_t \mid s_t)$.

*Definition 14 (**Multi-Armed Bandits (MAB)** [12]):* A K-armed bandit problem is defined as a tuple $\mathcal{M} = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a set of player's arm choices from K arms $a_t \in \mathcal{A} = \{a_1, ..., a_K\}$ at round t, and $\mathcal{R}$ is a set of outcome variables representing the rewards $r_t \in \{0, 1\}$.

Note that when unobserved confounders exist in the K-armed bandit, the model then would be established and replaced by $\mathcal{M} = (\mathcal{A}, \mathcal{R}, U)$, where U is the unobserved variable that implies the payout rate of arm a_t and the propensity score for choosing arm a_t [12].

*Definition 15 (**Contextual Bandits (CB)** [59]):* The contextual bandit is defined as a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ and $\mathcal{R}$ are the same notations as MAB. $\mathcal{X}$ is a set of contexts, i.e., the observed side information.

*Definition 16 (**Imitation Learning Model (IL)** [60]):* An imitation learning model is defined as a tuple $\mathcal{M} = (\mathcal{O}, \mathcal{T})$,

where $\mathcal{O}$ denotes the accessible high-dimensional observations $o \in \mathcal{O}$, while $\mathcal{T}$ denotes the trajectories generated by the expert policy $\mathcal{T} \sim \pi_D(\cdot \mid o)$.

*Definition 17 (**Dynamic Treatment Regime (DTR)** [61], [62]):* A dynamic treatment regime is defined as a sequence of decision rules $\{\pi_T : \forall T \in \mathbf{T}\}$, where $\mathbf{T}$ is a set of treatments. Each π_T is a mapping function from the values of treatments' and covariates' history H_T to the domain of probability distributions over T, denoted by $\pi_T(T \mid H_T)$.

● *Model-free methods*

Model-free RL methods typically have irreversible access to the world model, but learn the policy directly and purely from interactions with the environment, similar to the way we act in the real world [63]. Popular approaches include policy-based, value-based and actor-critic ones.

Policy-based methods directly learn the optimal policy π^* via the policy parameter θ to maximize the accumulated reward. They basically take the best of policy gradient theorem for deriving θ. Typical methods are Trust Region Policy Optimization (TRPO) [64], Proximal Policy Optimization (PPO) [65], etc., which use function approximation to adaptively or artificially adjust super-parameters to accelerate the methods' convergence.

In value-based methods, the agent updates the value function to gain the optimal one $Q^*(s, a)$, implicitly obtaining a policy. Q-learning, state–action–reward–state–action (sarsa) and Deep Q-learning Network (DQN) are typical value-based methods [1], [3], [4], [55]. The update rules for Q-learning and sarsa involve a learning rate α and a temporal difference error δ_t:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha \delta_t, \quad (11)$$

where $\delta_t = r_{t+1} + \gamma \arg\max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$ in off-policy Q-learning, and $\delta_t = r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$ in on-policy sarsa. However, it can only handle discrete spaces of states and actions. DQN with deep learning expresses the value or policy with a neural network, enabling to deal with continuous states or actions. It stabilizes the Q-function learning by experience replay and the frozen target network [3]. Improvements of DQN are Double DQN [66], Dueling DQN [67], etc.

Actor-critic methods combine the merits of networks from policy-based and value-based methods, where the actor network originates from the policy-based methods while the critic network from the value-based ones [68], [69]. The principle framework of actor-critics consists of two parts: i) actor: outputting the best action a_t based on a state s_t, which controls how the agent behaves with learning the optimal policy; ii) critic: computing the Q value of the action, which achieves evaluation of the policy. Typical methods include Advantage actor-critic (A2C) [69], Asynchronous Advantage actor-critic (A3C) [69], Soft actor-critic (SAC) [70], Deep Deterministic Policy Gradient (DDPG) [71], etc. Especially, SAC introduces a maximum entropy term to improve the agent's exploration and stability of the training process with the stochastic policy [70]; DDPG applies neural networks to operate on high-dimensional and visual state space. It embraces Deterministic Policy Gradient (DPG) [72] and DQN [4] methods to be roles

of the actor and critic, respectively, alleviating the high bias and high variance problems [71].

- *Model-based methods*

Without interacting with the environment directly, model-based RL methods majorly exploit the learned or given world model to simulate transitions, optimizing the target policy efficiently. It is similar to the way humans imagine in their mind [63]. We here introduce some common model-based RL algorithms according to how the model is used, i.e., the black-box model for trajectory sampling and white-box model for gradient propagation [73].

With an available black-box model, the direct way of applying it in policy learning is to plan in such a model. Methods include Monte Carlo (MC) [74], Probabilistic Ensembles with Trajectory Sampling (PETS) [75], Monte Carlo Tree Search (MCTS) [76]–[78], etc. MCTS is an extension sampling method of MC, and it basically adopts a tree search to determine the action with high probability to transit to the high-value state at each time step [76]. It has been applied in AlphaGo and AlphaGo Zero, to challenge the professional human players in the game of Go [77], [78]. On the other hand, an available model could be used to generate simulated samples to accelerate the policy learning or value approximation, and this process is well-known as the Dyna-style methods [79]. That is, models in the Dyna play a role of experience generator to produce augmented data. For instance, Model Ensemble Trust Region Policy Optimization (ME-TRPO) [80] learns a set of dynamic models with the collected data, with which it generates imagined experiences. Thereafter it updates the policy with augmented data in the ensemble model, using TRPO model-free algorithm [64]; Model-Based Policy Optimization (MBPO) [81] samples the branched rollouts with the policy and the learned model, and utilizes SAC [70] to further learn the optimal policy with augmented data.

With a white-box dynamics model, one can employ the internal structure to generate gradients of the dynamics so as to facilitate the policy learning. Typical methods include Guided Policy Search (GPS) [82], Probabilistic Inference for Learning Control (PILCO) [83], etc. GPS [82] utilizes the path optimization technique to guide the training process, resulting in improved efficiency. It draws samples via the Iterative Linear Quadratic Regulator (iLQR), which are used to initialize the neural network policy and further update the policy [73]. PILCO [83] assumes the dynamic model as the Gaussian process. After learning such a dynamic model, it evaluates the policy with approximate inference and acquires the policy gradients for policy improvement. In real-world applications, offline RL often matters, where the agent has to learn a satisfactory policy only from an offline experience dataset but with no interaction with the environment. One key challenge of offline RL belongs to the distributional shift problem, resulting from the discrepancy between the behavior policy for the training data and the current learning policy [73]. To bypass the distributional shift problem, [84] proposed a Model-Based Offline Policy Optimization (MOPO) algorithm. They derive a policy value lower bound based on the learned model and attempt to penalize the rewards by the uncertainty of the dynamics.

III. CAUSAL REINFORCEMENT LEARNING

Since causality and reinforcement learning can be umbilically tied with each other, it is desirable to investigate how they are effectively integrated to implement policy learning tasks or(and) causality tasks [4]. And such an implementation is called causal reinforcement learning, which is defined formally as follows.

Definition 18 (Causal Reinforcement Learning, CRL): CRL is a suite of algorithms which aim at embedding causal knowledge into RL for more efficient and effective model learning, policy evaluation, or policy optimization. It is formalized as a tuple $(\mathcal{M}, \mathcal{G})$, where $\mathcal{M}$ represents an RL model setting, e.g., MDP, POMDP, MAB, etc., while $\mathcal{G}$ stands for the causal-based information regarding an environment or task, e.g., causal structure, causal representation or features, latent confounders, etc.

Based on whether the causal information is empirically given or not, methods of causal reinforcement learning roughly fall into two categories, i.e., (i) methods based on the given or assumed causal information; and (ii) methods based on the unknown causal information which has to be learned by techniques. Causal information includes especially the causal structure, and causal representation or causal features, latent confounders, etc.

A. CRL with Prior Causal Information

Here we review CRL methods where the causal information is known (or given a prior) explicitly or implicitly. Generally, such methods assume the causal structure regarding to the environment or task is given a prior from the experts, which may include how many latent factors are, where the latent confounders locate or how they affect other observed variables of interest. As for the latent confounders scenarios, most of them would consider to eliminate the confounding bias towards the policy via proper techniques, while they use RL methods to learn the optimal policy. They may also theoretically prove the worst-case bounds of policy to achieve policy evaluation. As for the unconfounded scenarios, they use prior causal knowledge for sample efficiency, causal explanation or generalization in policy learning. During or before the modern RL procedure, these methods perform data augmentation with causal mechanisms; narrow the search space with causal information; or give preference towards those having causal influence.

We summarize these CRL methods according to different model settings, i.e., MDP, POMDP, bandits, DTR and IL[5].

1) MDP: Giving a digital advertising example, Li et al., [52] showed the existence of reinforcement bias in policy, which could be amplifiable in interaction with the environment. To correct the bias theoretically and practically, they

---

[4]Please see Appendix for preliminary of causality and RL.

[5]Please note that since there are close connections between these model settings, it is not restricted for a CRL method to strictly belong to one of these models only.

Models	Algorithms
MDP	IVOPE [85], IV-SGD and IV-Q-Learning [52], IVVI [86], CausalDyna [87], CTRL$_g$ and CTRL$_p$ [88], IAEM [89], MORMAXC [90], DOVI [91], FQE [92], COPE [93], off_policy_confounding [94], RIA [95], etc.
POMDP	CF-GPS [96], Gumbel-Max SCMs [97], CFPT [98], Decoupled POMDPs [99], PCI [100], partial history weighting [101], COMA [102], influence MOA [103], CCM [104], Confounded-POMDP [105], etc.
Bandits	Causal TS [12], DFPV [53], TS^{RDC*} [106], SRIS [107], Causal Bandit [108], [109], C-UCB and C-TS [110], UCB+IPSW [111], B-kl-UCB and B-TS [112], POMIS [113], [114], Unc_CCB [115], OptZ [116], CRLogit [117], [118], etc.
DTR	OFU-DTR and PS-DTR [62], UC^c-DTR [119], CAUSAL-TS* [120], IV-optimal and IV-improved DTRs [121], etc.
IL	CI [122], exact linear transfer method [123], Sequential π-Backdoor [124], CTS [125], DoubIL and ResiduIL [126], MDCE-IRL and MACE-IRL [127], etc.

proposed a class of instrumental variable-based RL methods under the two-timescale stochastic approximation framework, incorporating a stochastic gradient descent routine, and Q-learning algorithm, etc. They considered the MDP environment where noises may be correlated with states or actions, based on which they constructed the IVs. IVs prescribed state-dependent decisions. They used instrumental variables (IVs) to learn the causal effect of a policy, based on the given causal structure, and hereafter true rewards were debiased and the optimal policy was learned. Focusing on the case [128] where the confounder u_t at time t affected both the action a_t and the state s_{t+1} as illustrated in Fig. 3, Liao et al., [86] constructed a Confounded MDP model taking IVs and unobserved confounders (UCs) into consideration, namely CMDP-IV. They derived a conditional moment restriction, by which they identified the confounded nonlinear transition dynamics under the additive UCs assumption. With the primal-dual formalization of such a conditional moment restriction, they eventually proposed an IV-aided Value Iteration (IVVI) algorithm to learn the optimal policy in offline RL. Zhang et al., [90] investigated the problem of estimating the MDP with the presence of unobserved confounders. If ignoring such confounders, sub-optimal policies may be obtained. Hence, they utilized causal languages to formalize the problem, and demonstrated explicitly two classes of policies (i.e., the experimental policy and the counterfactual one). After proving the superiority of the counterfactual policy instead of the experimental policy, they restricted standard MDP methods to search in the space of counterfactual policies, with empowered performance in speed and convergence than state-of-the-arts algorithms. Wang et al., [91] studied the problem of constructing the information gains from the offline dataset so as to improve the sample efficiency in the online setting. Hence, they proposed a Deconfounded Optimistic Value Iteration algorithm (DOVI). As shown in Fig. 4, in the offline setting, they assumed the confounders were partially observed, whose confounding bias could be corrected by the

backdoor criterion; while in the online setting, they assumed the confounders were unobserved, whose confounding bias could be corrected by intermediate states via the frontdoor adjustments. They eventually gave the regret analysis of their proposals.

To handle the issue of data scarcity and mechanism heterogeneity, Lu et al., [88] proposed a sample-efficient RL algorithm that leveraged SCMs to model the state dynamics process, and aimed for the general policy over the population as well as the personalized policy over each individual. In particular, as for the general policy, based on the causal structure in a MDP, they assumed the state variable s_{t+1} satisfies the SCM,

$$s_{t+1} = f(s_t, a_t, u_{t+1}), \tag{12}$$

where f is a function representing the causal mechanism from the causes to s_{t+1}, a_t is the action at time t, and u_{t+1} stands for the noise term of s_{t+1}. As for the personalized policy for an individual, they assumed the state variable s_{t+1} satisfies the SCM,

$$s_{t+1} = f(s_t, a_t, \theta_c, u_{t+1}), \tag{13}$$

where f represents the overall family of causal mechanisms, and θ_c captures change factors that may vary cross individuals. With these two state transition processes, they then employed Bidirectional Conditional generative adversarial network framework to estimate f, u_{t+1} and θ_c. With performing the counterfactual reasoning, the data scarcity problem was mitigated. Zhu et al., [87] adopted the same idea of counterfactual reasoning based on SCMs to improve the sample efficiency. They modeled the time-invariant property (e.g., the mass of an object in robotic manipulations) as an observed confounder affecting states variables at all time steps, and proposed a Dyna-style causal RL algorithm. Following the goal-conditioned MDP model [129], Feliciano et al., [130] accelerated the exploration and exploitation learning processes to learn policies via causal knowledge. Assuming the presence of a possibly incomplete or partially correct casual graph, they narrowed the search space of actions via graph queries, and developed a method to guide the action selection policy. To efficiently perform RL tasks, Lu et al., [131] studied causal MDPs, where interventions and states/rewards constituted a three-layer causal graph. The states or rewards (outcomes) was on the top layer, parents of the outcome were on the middle layer, while the direct interventions (manipulable) were on the bottom layer. Then given such prior causal knowledge, they proposed the causal upper confidence bound value iteration (C-UCBVI) algorithm and causal factored upper confidence bound (CF-UCBVI) algorithm, to circumvent the curse of dimensionality of actions or states. They finally proved the regret bounds for verification.

As for accelerating the explanation for decisions of the system, Madumal et al., [132] employed causal models to derive causal explanations of the behavior in model-free RL agents. Specifically, they introduced an *action influence model* where the variables in the structural causal graph were not only states but also actions. They assumed that a DAG with causal directions was given a prior, and presented an approach
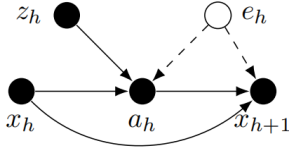
Fig. 3. An example of causal graphical illustration of confounded MDP, where $\{z_h\}$ is a sequence of IVs, $\{e_h\}$ is a sequence of unobserved confounders, and x_h is the current state variable at the hth time step [86].

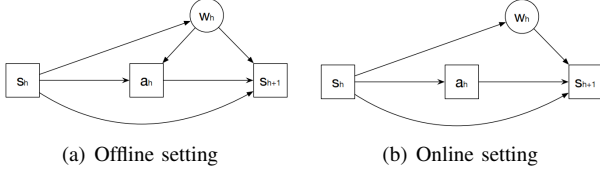(a) Offline setting   (b) Online setting

Fig. 4. An example of causal graphical illustration of confounded MDP in the offline and online settings, where $\{w_h\}$ is a sequence of unobserved confounders at the hth time step [91].

to learn such a structural causal model so as to generate contrastive explanations for why and why not questions. Finally, they were enable to explain why new events happened by counterfactual reasoning. As for the data efficiency and generalization problems, Zhu et al., [89] introduced an invariance of action effects among states, which was inspired by the fact that the same action could have similar effects among different states on the transitions. Utilizing such an invariance, they proposed a dynamics-based method termed as IAEM (Invariant Action Effect Model) for generalization. Specifically, they first characterized the invariance as the residual between representations from neighboring states. And they applied a contrastive-based loss and a self-adapted weighting strategy for better estimation of these invariance representations. Guo et al., [95] proposed a relation intervention MBRL method to generalize to dynamics unseen environments, where a latent factor that changed across environments was introduced to characterize the shift of the transition dynamics. Such a factor was drawn from the history transition segments. Given the causal graph, they introduced interventional prediction module and a relational head to reduce the redundant information from the factors.

Off-Policy Evaluation (OPE) of sequential decision-making policies under uncertainty is a fundamental and essential problem in batch RL. However, there existed challenges which hammered its development, including the confounded data, etc. The presence of confounders in data would render the evaluation of new policies unidentifiable. To handle such an issue, Kallus et al., [128] explored the partial identification for off-policy evaluation in the setting of infinite-horizon RL with confounders. Specifically, they first assumed the stationary distributions for unobserved confounding and the subjection to a sensitivity model for the data. In their satisfied causal model, the confounder u_t at the timestep t affected both the action a_t and the state s_{t+1}. They computed the sharpest bound of the policy value by characterizing and optimizing the partially identified set. They ultimately proved their proposed approximation method was consistent to establish the bounds

in practice. In the cases where the observed action a_t, state s_{t+1} and reward r_t were confounded by the i.i.d. observed variable u_t at each time step t, Bennett et al., [133] proposed an OPE method in the infinite-horizon setting, to estimate the stationary distribution ratio of states in the presence of unobserved confounders and to estimate the policy value while avoiding direct modeling of reward functions. They proved the identifiability of the confounding model, They also gave statistical consistency and provided error bounds under some assumptions. When encountering unobserved confounding, Namkoong et al., [94] analyzed the sensitivity of OPE methods in sequential decisions making cases, and demonstrated that even a small amount of confounding could induce heavily bias. Hence, they proposed a framework to quantify the impact of the unobserved confounding and to compute the worst-case bounds, which restricted the confounding in a single time step, i.e., the confounders might only directly affect one of the many decisions and further influence future rewards or actions. The bounds on the expected cumulative rewards were estimated by an efficient loss-minimization-based optimization, with the statistical consistency. Both previous methods [94], [128] adopted importance sampling approaches to compute the worst-case bounds of a new policy. While Kallus et al., [128] constructed bounds in the infinite horizon case and Namkoong et al., [94] deduced the worst-case bounds with restricted confounding at a single time step, they however could not render the finite case with confounding at each step tractable. To conquer this challenge, Bruns [92] developed a model-based robust MDP approach to calculate finite horizon lower bounds with confounders that were drawn i.i.d. each period, combining the sensitivity analysis. For the cases where the confounders were persistently available, off-policy evaluation was far more challenging. Recent methods for OPE in the MDP or POMDP models [94], [99], [105], [128], [133], [134] did not considered the case which aimed at estimating confidence intervals offline for the target policy's value in infinite horizons. Shi et al., [93] took great care of such a case. They modeled the data generation process based on a confounded MDP additionally with some observed immediate variables, and named this causal diagram as CMDPWM (Confounded MDP With Mediators). With the mediators, the target policy's value was proved to be identifiable. And they provided an algorithm to estimate the off-policy values robustly. Their method helped ride-sharing companies solve the problem of evaluating different customer recommendation programs that may contained latent confounders. If ignoring confounders when minimizing the Bellman error, one might obtain biased Q-function estimates. Chen et al., [51] exploited both IVs and RL in the language of OPE, proposing a class of IV methods to overcome these confounders and achieve identifying the value of a policy. They conducted experiments based on the MDP environments with a set of OPE benchmark problems and various IV comparisons.

2) POMDP: To shrink the mismatch between the model and the true environment as well as improve sample efficiency, Buesing et al., [96] leveraged counterfactual reasoning and proposed the Counterfactually-Guided Policy Search (CF-GPS) algorithm for learning policies in POMDPs from off-
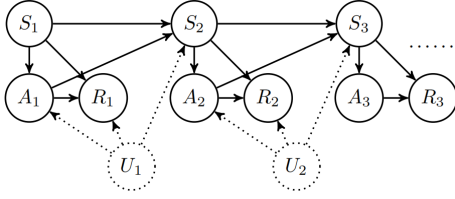
Fig. 5. An example of causal graphical illustration of MDP with unmeasured confounders, where $U.$ are confounders. The action policy, state transition and reward emission are all confounded at every step [133].
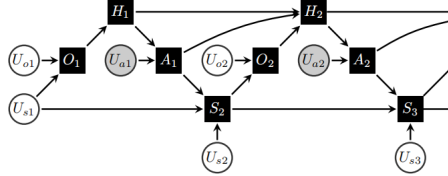

Fig. 6. An example of causal graphical illustration of POMDP, where $U.$ are noise variables, and H_t are histories. The mechanism that generates the actions A_t relies on the histories H_t [96].

(a) Under logging policy
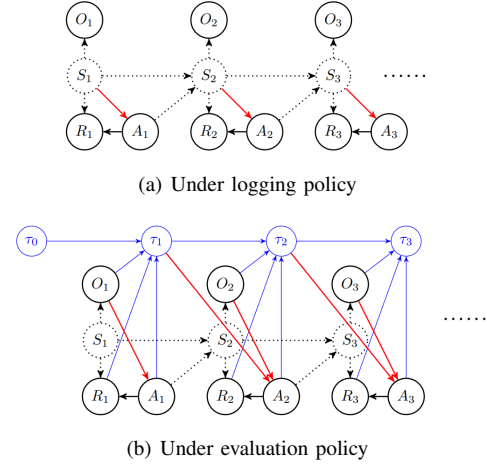

(b) Under evaluation policy

Fig. 7. Examples of causal graphical illustrations of POMDP under the logging policy and the evaluation policy, where the red arrows denote the dependence relations between the policy actions A_t and the states S_t in (a) (while in (b) the red arrows between the policy actions A_t and the observations O_t and the previous trajectories τ_t). Blue arrows represent the dependence relations of the observed trajectories [100]. The randomized given logging policy with POMDP generates the true trajectories while the evaluation policy denotes the deterministic target policy that needs evaluated.

policy experience. They first modeled POMDPs using the SCM representation as demonstrated in Fig. 6, and showed how to use the given structural causal models to counterfactually evaluate the arbitrary policies on individual off-policy episodes. Further, with available logged experience data, they made generalization to the policy search case via the counterfactual analysis in SCMs. Here they assumed the transition function is deterministic. To extend the counterfactual reasoning from deterministic transition functions to stochastic transitions, Oberst et al., [97] made further assumptions of a specific SCM, that is, the Gumbel-Max SCM, and proposed an off-policy evaluation method via counterfactual queries in categorical distributions. In particular, such a given class of SCM enabled to generate counterfactual trajectories in finite POMDPs. They identified those episodes where the counterfactual difference in reward is most dramatic, which could be used to flag individual trajectories for review by domain experts. Given similarly as [97] that the POMDP followed a particular Gumbel-Max SCM, Killian et al., [98] presented a Counterfactually Guided Policy Transfer algorithm to fulfill the policy transfer from the source domain to a target domain offline, off-policy clinical settings. They followed the identical model formalization as [96], POMDP with SCMs. Intuitively, they leveraged the observed transition probability and the trained treatment policy in the source domain to aid guiding the counterfactual sampling in the target domain. Instead of using counterfactual analysis, Tennenholtz et al., [99] took the best of the effects of intervention on the world with a different action and an evaluating policy, and used them to fulfill the off-policy evaluation of sequential decisions. Their work was majorly based on the POMDPs, and especially, the Decoupled POMDPs where state variables were partitioned into two disjoint set, the observed and the unobserved ones. They showed that such decoupled new models could reduce the estimation bias from the general POMDPs. However, their method was only valid in tabular settings where states and

actions were discrete.

To generalize the work [99] to a non-tabular setting, Bennett et al., and Shi et al., [100], [105] also implemented the OPE task under the POMDP model, i.e., evaluating a target policy in the POMDP given the observational data generated by some behavior policy. Bennett et al., [100] calibrated the proximal causal inference (PCI) approach to the sequential decision making setting. In particular, they formalized graphical representations under POMDP model under the logging policy and the evaluation policy. Based on the graphs depicted in Fig. 7, they then provided the identification results of evaluating the target policy via PCI, under some assumptions about the independence constraint and bridge functions. They finally established an estimator for evaluation. On the other hand, Shi et al., [105] deduced an identification method for OPE in POMDP model in the presence of latent confounders, where the behavior policy depended on unobserved state variables while the evaluation policy only on observable observations. Their method was based on a weight bridge function and a value bridge function. Afterwards, they proposed minimax estimation methods to learn these two bridge functions, deploying function approximation techniques. They further proposed three estimators to learn the target policy's value, on the basis of the value function, the marginalized importance sampling and the doubly-robustness. Their proposal was suitable in settings with continuous or large observation/states spaces.

Nair et al., [134] also zeroed in on the OPE problem in POMDPs, same as [99]. But unlike previous works that relied on the invertibility of certain one-step moment matrices, they relaxed such an assumption from one-step proxies to multi-step one with the past and future, motivated by spectral learning. And they developed an Importance Sampling algorithm that depended on the rank conditions on probability matrices, and not on the sufficiency conditions of observable

trajectories. Their method was executed with various kinds of causal structures for proxies with the confounder. The above-mentioned work including [99], [100], [134] attempted to correct confounding bias in the context of POMDP model, whereas Hu et al., [101] worked in the cases with no confounding. They proposed an OPE method to identify causal effects of a target policy also in the POMDP model, by using partial history importance weighting technique. They further established upper and lower bounds on the error between the estimated and true values.

As for the cooperative multi-agent settings, previous methods were faced with decentralized policies learning and multi-agent credit assignment problems. To conquer such challenges, Foerster et al., [102] proposed a multi-agent actor-critic algorithm called COunterfactual Multi-Agent (COMA) policy gradient, based on the POMDP model along with one more element standing for the multiple agents. In particular, their COMA majorly enjoyed three merits. Firstly, they utilized a centralized critic to estimate a counterfactual advantage for decentralized policies. Secondly, they dealt with the credit assignment problem in virtue of a counterfacutal baseline, which enabled marginalizing out a single agent's action while keeping all the other agents' actions fixed. Finally, their critic representation allowed efficient estimation for the counterfactual baseline. Unlike [102] which resorted to centralized learning, Jaques et al., [103] explored a more realistic scheme where each agent was trained independently. Specifically, they [103] focused on reaching the goal of coordination and communication in the multi-agent RL setting. Their intuitive idea was to reward intrinsically to those agents who were having high causal influences on other agents' actions, where the causal influences were evaluated by counterfactual analysis. They conducted extensive experiments on two Sequential Social Dilemmas (SSDs), which were partially observed, spatially and temporally extended multi-agent games, and they showed the feasibility of their proposed social influence rewards with communication protocols. They also fostered to model other agents, allowing each agent to compute its social influence independently in a decentralized way, i.e., without access to other agents' reward functions. Barton et al., [104] measured the collaboration between agents via Convergent Cross Mapping (CCM), which examined the causal influence using predator agents' positions over time.

3) Bandits: As for the Multi-Armed Bandit (MAB) problem, Bareinboim et al., [12] first showed that previous bandit methods may fall into the sub-optimal policy under the existence of unobserved confounders. Hence, they proposed a new bandit problem with unobserved confounders using causal graphical representation, where the confounder affects both the action and the outcome. In this setting, an observed variables $X_t \in \{x_1..., x_k\}$ was encoded as the agent's arm choice from k arms, arms (or actions) were identified as interventions on the causal graph, and $Y_t \in \{0, 1\}$ was a reward (or outcome) variable from choosing the arm X_t under the unobserved confounder U_t. They hereafter developed an optimization metric employing both experimental and observational quan-
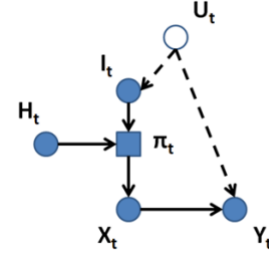

Fig. 8. An example of causal graphical illustration of MAB with unobserved confounders U_t, where I_t is the agent's intended arm choice at step t, π_t denotes the policy, X_t is the arm choice, Y_t is the Bernoulli reward and H_t denotes the agent's history [106].

tities and empowering Thompson Sampling [6], namely Causal Thompson Sampling method. Here they estimated the Effect of the Treatment on the Treated group (ETT). Motivated by the unobserved confounder pitfalls in MAB from [12], Forney et al., [106] employed a different vehicle for an online learning agent: counterfactual-based decision making. They first exploited counterfactual reasoning to generate counterfactual experiences by active agents. And they proposed a Thompson Sampling method where the agents combined observational, experimental and counterfactual data to learn the environments in the presence of latent confounders. It's worth noting that their confounding MAB model considered the agent's arm choice predilections and pay out rates at each round, as shown in Fig. 8. Following the similar causal bandit problem as [12], Lattimore et al., [109] formally introduced a new causal bandit framework which subsumes traditional bandits as well as the contextual stochastic bandits. In their framework, observations were obtained after performing each intervention and the agent could control the allowed actions. To improve the learning rate at which good interventions were learned, they proposed a general algorithm to estimate the optimal action, assuming the causal graph was given (but can be arbitrary). They also proved that the algorithm enjoyed the strictly better regret than those with no exploited causal information.

To majorly improve the performance over the method [109] in both theoretical and practical ways, Sen et al., [107] proposed an efficient successive rejects algorithm with importance sampling, also assuming the causal structure was given a prior (so did a small part of interaction strengths). Further, they provided the first gap dependent error and simple regret bounds to identify the best soft intervention at a source variable, so as to maximize the expected reward of the target variable. Since previous works only considered localized interventions [107], [109], i.e., the intervention propagated locally and only to the intervened variable's neighbors, Yabe et al., [108] introduced a causal bandit method via propagating inference, which achieved propagating throughout the causal graph with arbitrary interventions. In their method, the causal graph and a set of interventions were given. They also deduced the simple regret bound, which was logarithmic with respect to the number of arms.

Harnessing a given causal graph as background knowledge,

---

[6]Thompson Sampling (TS) is a traditional bandit algorithm.

Lu et al., [110] proposed two algorithms, namely causal upper confidence bound (C-UCB) and causal Thompson Sampling (C-TS), to learn the optimal interventions sequentially for the bandit problem. They assumed that the distribution of parents of the reward variable for each intervention is known. Their proposals, including the linearity cases, aimed to reduce the amount of needed explorations and enjoyed improved cumulative regret bounds. Since situations exist where performing the intervention is costly, Nair et al., [135] considered causal bandits in a given causal graph with two settings: with and without budget constraints. Such constraints determined whether to take the trade-off between costly interventions and economical observations. Hence, they proposed two algorithms, γ-NB-ALG and CRM-NB-ALG that minimized the expected simple regret and the expected cumulative regret respectively with the trade-off, utilizing causal information; and an algorithm C-UCB-2 without the budget constraint in general graphs, making the same assumption as [110] for the distribution of reward's each parent.

Lee et al., [113] uncovered a phenomenon that traditional strategies for MAB, including intervening on all variables at once or on all subsets of variables, would lead to suboptimal policies. Hence, to guide the agent whether conducting a causal intervention or not, and which variables to intervene, they first employed the language of SCMs to characterize a new type of Multi-Armed Bandit (MAB), called SCM-MABs. They then investigated the structural properties of a SCM-MAB, which could be computed by any arbitrary causal model. Furthermore, they introduced an algorithm to identify a special set of variables, which tells the possibly-optimal minimal intervention set for the agent, allowing the existence of unobserved confounders. Since in many real-world applications, not all observed variables are manipulable, Lee et al., [114] studied such a setting of SCM-MAB problems and proposed an algorithm to identify the possibly-optimal arms. Their method was an enhanced version of the MAB algorithm, based on the crucial properties in the causal graph with manipulability constraints. Lee et al., [136] zeroed in on the mixed policies which consisted of a collection of decision rules where each rule was assigned with different actions for interventions given the contexts. In particular, they first characterized the non-redundancy for the mixed polices, with a graphical criterion designed to identify the unnecessary contexts for the actions. Further they offered sufficient conditions to characterize the non-redundancy under optimality, which fostered an agent to adjust its suboptimal policy. Such characterizations lied a stark benefit to fundamentally understand the space of mixed policies and refine the agent's strategy, helping the agent fast converge to the optimum.

As for the contextual bandit problem, Subramanian et al., [115] formalized a nuanced contextual bandit setting where the agent had the choice to perform either a targeted intervention or a standard interaction during the training phase and had access to the causal graph information. This is the first paper to introduce targeted interventions in the contextual bandit problem, and the work to integrate the causal-side information (namely causal contextual bandit problems) [115]. Their
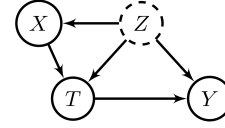

Fig. 9. An example of causal graphical illustration of contextual MAB, where X denotes observed covariates, T is an observed treatment, Y is an outcome and Z stands for latent confounders [116].

proposed algorithm was based on minimizing an entropy-like measure, which reduced the demand for a large number of samples. To embrace the merits of offline causal inference and online contextual bandit learning, Li et al., [111] proposed a framework to select some logged data to improve the accuracy of online decision making problems, including context-independent and context-dependent decisions. Given the contextual bandit model with confounders, they also derived the upper regret bounds of their forest-based algorithms. Instead of dealing with the online learning problem, Zhang et al., [112] studied the offline transfer learning problem in RL scenarios, leveraging causal inference and allowing the existence of unobserved confounders. In particular, given a causal model, they first connected algorithms to achieve identifying causal effects (from trajectories of the source agent to the target agent's action) and off-policy evaluation. In the cases where causal effects could not be identified, they proposed a new strategy to extract causal bounds, which were used to learn the policy with high learning speed.

As for the policy evaluation problem with latent confounders, Bennett et al., [116] developed an importance weighting approach based on the contextual decision making setting. Given the DAG representation of model where the confounder affected the proxies, treatment and the outcome in Fig. 9, they induced an adversarial objective for selecting optimal weights with some given function classes. Such optimal weights were proved to produce consistent policy evaluations. This OPE work with latent confounders in contextual bandit settings was considered as a special case of [133]. Given the identical causal graph as [116] in Fig. 9 where the confounder affected the proxies, treatment and the outcome, Xu et al., [53] proposed a Deep Feature Proxy Variable method (DFPV) to estimate the causal effect of treatment on outcomes with unobserved confounders, using proxy variables of confounding. Their proposal, which was based on the proximal causal inference (PCI), was applied to bandit OPE with continuous actions. Note that previous work [99] applied PCI as well for OPE, but with discrete action-state spaces.

Gonzalez et al., [137] proposed a method for the agent to make decisions in an uncertain environment which was governed by causal mechanisms. Their method allowed the agent to hold beliefs about the causal environment, which were kept updating using interactive outcomes. In the experiments, they assumed the agent knew the causal structure of the environment.

As for the policy improvement problem with unobserved confounders, where the confounder affected both treatments and outcomes, Kallus et al., [117] developed a robust framework to optimize the minimax regret of a candidate policy
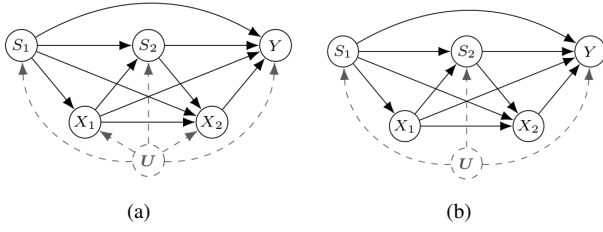
Fig. 10. Examples of causal graphical illustrations of (a) a DTR with 2 stages of intervention; and (b) a DTR in (a) under sequential interventions towards X_1 and X_2: $do(X_1 \sim \pi_1(X_1|S_1), X_2 \sim P(X_2|S_1, S_2, X_1))$. **X** denotes the set of treatments, **S** denotes the set of covariates, Y is the outcome, and U is the unmeasured confounder [119].

against a baseline policy, motivated by sensitivity analysis in causal inference. They optimized over the uncertainty sets with nominal propensity weights, using subgradient descent method to perform policy optimization. Afterwards, Kallus et al., [118] provided an extended version of the above method [117], which induced sharp regret bounds in binary- as well as multiple-treatment settings. Further, they generalized theoretical guarantees on the performance of the learned policy, on the basis of previous theories [117].

4) DTR: Dynamic Treatment Regime (DTR) aims at offering a set of decision rules, one per stage of intervention, that indicate what treatments are for an individual patient to improve the long-term clinical outcome, given the patient's evolving conditions and histories. DTR differs from the traditional MDP in that it does not require the Markov assumption and what it estimates is an optimal adaptive dynamic policy that makes its decision based on all previous information available [86]. It is an attractive vehicle for personalized management in medical research [119].

Apart from Definition 17, DTR could also be defined in detail with the language of SCM in causality [61]. That is, a DTR is defined as a SCM, i.e., a tuple $< \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{u}) >$, where **U** represents exogenous unobserved variables (or noise variables); $\mathbf{V} = \{\mathbf{X}, \mathbf{S}, Y\}$ is a set of endogenous observed variables, **X** is a set of action variables, **S** a set of state variables, and Y is the outcome; **F** represents a set of structural functions; and $P(\mathbf{u})$ is the probability distribution of noises **u**. For stage $k = 1, ..., K$, a state is determined by a transition function: $s_k = \tau_k(\bar{\mathbf{x}}_{k-1}, \bar{\mathbf{s}}_{k-1}, \mathbf{u})$, a decision is decided by a behavior policy: $x_k = f_k(\bar{\mathbf{s}}_k, \bar{\mathbf{x}}_{k-1}, \mathbf{u})$, and an outcome ranging from 0 to 1 is determined by a reward function: $y = r(\bar{\mathbf{x}}_K, \bar{\mathbf{s}}_K, \mathbf{u})$. $\bar{\mathbf{x}}_k$ stands for a sequence of values from variables $\{X_1, ..., X_k\}$, i.e., $\bar{\mathbf{x}}_k = \{x_1, ..., x_k\}$ [61]. Especially, when there is only one stage in DTR with an empty set of states, the model of such a DTR is a multi-armed bandit model.

To learn the optimal DTR in the online RL setting, Zhang et al., [119] developed an adaptive RL algorithm with the near-optimal regret. Based on the DTR structure where the confounders existed as demonstrated in Fig. 10, they derived causal bounds about the dynamics of system, which were integrated to calibrate the algorithm with accelerated and improved performance. Their bounds were relative to the domains of treatments **X** and covariates **S**.

If the domains were huge, such regrets would not be

attainable. Hence, to reduce the dimensionality of the policy space, Zhang et al., [62] first proposed an algorithm to remove irrelevant treatments or evidence by exploiting the independence constraints in the causal structure of the DTR. Then, they designed two online RL algorithms to identify the optimal DTR in an unknown environment, by making the best of the causal order relationships in the form of a given causal diagram. One algorithm employed the principle of optimism in the face of uncertainty to deduce tight regret bounds for DTR while the other one leveraged the heuristic of posterior sampling.

Zhang et al., [120] took care of the mixed policy scopes, which consisted of a collection of treatment regimens resulted from all possible combinations of treatments. Specifically, they designed an online learning algorithm to identify an optimal policy with mixed policy scopes, achieving sublinear cumulative regret. Further, provided with a causal diagram, they introduced a parametrization for SCMs with finite latent states, which enabled them to parametrize all observational and interventional distributions through using the minimal number of decomposition factors, named c-components. They then proposed another practical policy learning algorithm based on the Thompson sampling, which was computationally feasible and enjoyed the merit of asymptotic similar bound on the cumulative regret.

Chen et al., [121] considered the problems of estimating and improving DTR with a time-varying instrumental variable (IV) in the case where there existed unobserved confounders affected the treatment and the outcome. They thus developed a generic identification framework, under which they derived a Bellman equation based on the classical Q-learning and defined a class of IV-optimal DTRs. They further proposed an IV-improvement framework, which enabled to strictly improve the baseline DTR under the multi-stage setting.

5) IL: Imitation learning (IL) is originated from the fact that children learn skills via mimicking adults, which therefore aims at learning imitating policies from demonstrations generated by an expert [138], [139]. Existing IL methods roughly fall into two categories, namely behavior cloning (i.e., directly mimicking the behavior policy of an expert) and inverse reinforcement learning (i.e., learn a reward function from expert's trajectories and apply RL method for the policy). Here, we start to survey some IL methods along with causality.

To endow the agent to obtain imitability to learn an optimal policy, Zhang et al., [122] started with SCMs and utilized causal languages to propose a method to learn an imitating policy from combinations of demonstration data, allowing the existence of unobserved confounders. Their method took a causal diagram, a policy space and an observational distribution as inputs, with a complete graphical criterion for determining the imitation feasibility. The demonstrator and the learner might have different observations. Similarly, Etesami et al., [123] focused on addressing the differences and shifts between (a) the demonstrator's sensors, (b) our sensors and (c) the agent's sensors. According to the known causal graphs between states, actions, outcomes and observations in source and target domains as depicted in Fig. 11, they developed algorithms to identify effects of the demonstrator's actions
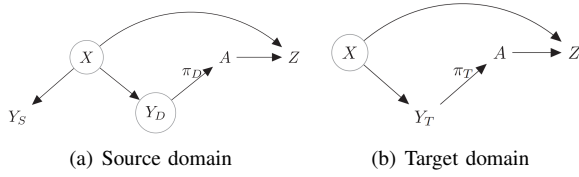
Fig. 11. Examples of causal graphical illustrations of IL in the (a) source domain; and (b) target domain. X denotes the state, A denotes the action, and Z is the outcome. Y_D stands for the demonstrator's input, generated by the demonstrator's sensors, while Y_S the spectator's observation of the state of the source system. Y_T denotes in the target domain the input from the target agent's sensors [123].

Fig. 12. An example of causal graphical illustrations of IL that captures the temporally correlated noises, where such noises u_t are the latent confounding variables that induce spurious correlations between states s_t and actions a_t [126].

and do imitation learning. They also provided theoretical guarantees for the estimated policy. Tennenholtz et al., [125] handled the problem of using expert data with unobserved confounders and possible covariate shifts for imitation learning and reinforcement learning. They first defined a contextual MDP setup, which implied its underlying causal structure. In the IL setting (without access to rewards), they claimed that in the case where the reward depends on the context, ignoring the covariate shifts would induce catastrophic errors and render non-identifiability of model. In the RL setting, they proved that when the reward is provided, the optimality could be reached with the expert data and even arbitrary hidden covariate shifts. Eventually they proposed an algorithm based on the corrective trajectory sampling with convergence guarantees.

The above solutions [122], [123], [125] however were limited by the single-stage decision-making setting. To explore the mismatch case in sequential settings where the learner must take multiple actions per episode, Kumor et al., [124] first induced a sufficient and necessary graphical criterion for determining the imitation feasibility, based on a known causal structure with domain knowledge of the covariates and targets. Thereafter, they proposed an efficient algorithm to optimize the policy for each action, while determining the imitability. They also proved the completeness of the criterion and verified the favorable performance of their method.

Empirically, we might encounter the data which were corrupted by temporally correlated noises, producing spurious correlations between states and actions and resulting in unsatisfactory policy performance, as illustrated in Fig. 12. To tackle this issue, Swamy et al., [126] leveraged variants of instrumental variable regression methods, taking the past states as instruments, and presented two algorithms to eliminate the spurious correlations. Dealing with the confounding problem based on the SCM, in IL, their methods derived consistent policies.

Compared with traditional IRL methods, those based on the maximum causal entropy enjoyed some merits, such as less specific assumptions on the expert behavior or good performances with a wide variety of reward functions, etc [140], [141]. Such a causal entropy allows to describe some structured qualities for the policy. However, these methods have to restrict to the finite horizon setting. To this end, Zhou et al., [127] extended the maximum causal entropy framework to infinite time horizon setting. They considered the maximum discounted causal entropy as well as the maximum average
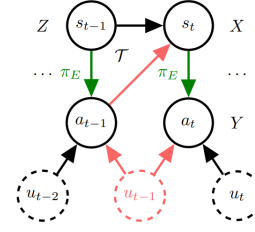
causal entropy, and proposed a gradient-based algorithm with the stationary soft Bellman policy.

B. CRL with Unknown Causal Information

In this subsection, we take a review on CRL methods where the causal information is unknown and has to be learned in advance. Compared with the first category, it is more challenging. And methods usually contain two phases: causal information learning and policy learning, which are carried out iteratively or in turn. Regarding the policy learning phase with causality, the paradigm is consistent with that of the first category. Regarding the causal information learning, it usually involves the causal structure learning (that may include latent confounders, or causes of actions or rewards), or causal representation/feature/abstraction learning. Most of these CRL methods adopt casual discovery techniques (including variants of PC [14], CD-NOD [34], Granger [41], etc), intervention, deep neural networks (including Variational Auto-Encoder [142], etc) for causal information learning.

We summarize these methods according to different model settings, i.e., MDP, POMDP, bandits and IL.

TABLE II
PART OF CRL ALGORITHMS WITH UNKNOWN CAUSAL INFORMATION

Models	Algorithms
MDP	ICIN [129], FOCUS [143], causal InfoGAN [144], NIE [145], DBC [146], IBIT [147], CoDA [148], CAI [149], FANS-RL [150], CDHRL [151], CDL [152], etc.
POMDP	causal curiosity [153], Deconfounded AC [154], RL with ASR [155], AdaRL [156], causal states [157], IPO [158], etc.
Bandits	CN-UCB [159], Multi-environment Contextual Bandits [160], Linear Contextual Bandits [161], etc.
IL	CIM [60], CPGs [162], causal misidentification [163], CIRL [164], CILRS [165], ICIL [166], cause-effect IL [167], copycat [168]–[171], etc.

1) MDP: To show that a causal world-model can surpass a plain world-model for the offline RL in both theoretical and practical ways, Zhu et al., [143] first proved generalization error bounds of the model prediction and policy evaluation with causal and plain world-models, and proposed an offline algorithm called FOCUS based on the MDP. FOCUS aimed at learning the causal structure from data by the extended PC causal discovery methods, and leveraging such a learned

structure to optimize a policy via an offline MBRL algorithm, MOPO [84].

Méndez-Molina [172] presented in his research proposal that there existed certain characteristics in robotics that accelerated causal discovery via interventions. Hence, he aimed at providing an intelligent agent that could simultaneously identify causal relationships as well as learn policies. Based on MDPs, he took inspiration of the Dyna-Q algorithm, and after discovering the underlying causal structure as the model, he used it to learn a policy for a given task by a model-free method. To capture causal knowledge in the RL setting, Herlau et al., [145] introduced a causal variable in the DOORKEY environment, which was served as a manipulatable mediating variable to predict the reward from the policy choice. Such a causal variable was binary, and was identified by maximizing the natural indirect effect (NIE) in mediation analysis. Thereafter, they built a small causal graph between this causal variable, policy choice and the return so as to help optimize the policy. To enable the agents to complete goal-conditioned tasks with causal reasoning capability, Nair et al., [129] formalized a two-phase meta-learning-based algorithm. In the first phase, they trained causal induction models from visual observations, which induced a causal structure via an agent's interventions; in the second phase, they leveraged the induced causal structure for a goal-conditioned policy with an attention-based graph encoding. Their method improved generalization abilities towards unseen causal relations as well as new tasks in new environments. Aiming at goal-directed visual plans, Kurutach et al., [144] combined the generalization merit from deep learning of dynamics models with the effective representation reasoning from classical planners, and developed a causal InfoGAN framework to learn a generative model of high-dimensional sequential observations. Via such a framework, they obtained a low-dimensional representation which stood for the causal nature of data. With this structured representation, they employed planning algorithms for goal-directed trajectories, which were then transformed into a sequence of feasible observations. Ding et al., [173] augmented Goal-Conditioned RL with Causal Graph (CG), that characterized causal relations between objects and events. They first used interventional data to estimate the posterior of CG; and then used CG to learn generalizable models and interpretable policies.

For complex tasks with sparse rewards and large state spaces, it is hard for Hierarchical Reinforcement Learning (HRL) to discover high-level hierarchical structures to guide the policy learning, due to the low exploration efficiency. To leverage the advantages from causality, Peng et al., [151] proposed a Causality-Driven Hierarchical Reinforcement Learning (CDHRL) framework to achieve effective subgoal hierarchy structures learning. CDHRL contains two processes: causal discovery and hierarchy structures learning, which are boosted by each other. Building discrete environment variables in advance in the subgoal-based MDP model, they adopted an iterative way to progressively discover the causal structure between such environment variables, and construct the hierarchical structure where reachable subgoals are those controllable environment variables from the discovered causal-
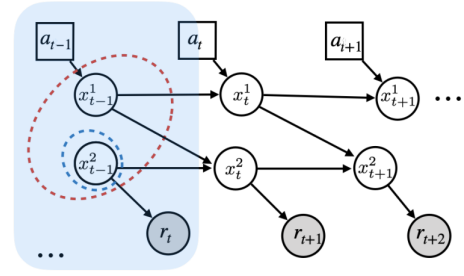

Fig. 13. An example of causal graphical illustration of MDP, with states $s = \{x^1, x^2\}$ when ignoring timesteps [177]. Note that x^2 is the only parent of the reward r.

ity. They evaluated their framework's efficiency in both causal discovery and hierarchy construction in complex tasks, i.e., 2d-Minecraft [174] and simplified sandbox survival games Eden [175].

It is hard for traditional methods to learn the minimal causal models of many environments. Furthermore, causal relationships in most environments may vary at every time step, termed as the time-variance problem [176]. To address such a problem, Luczkow [176] introduced a local representation (f, g) in an MDP that satisfies

$$s_{t+1} = f(s_t, a_t, g(s_t, a_t)), \qquad (14)$$

where f is a function representing the causal mechanism from the causes to the next state s_{t+1}, a_t is the action at time t, and $g(\cdot)$ is a function predicting the causal structure between states at each time step. Following this model formalization, they accomplished learning both causal models at the time step and the transition function, which could be used to planning algorithms and produced a policy. To extend existing stationary frameworks [156] to the non-stationarity settings where transitions and rewards could be time-varying within an episode or cross episodes in environments, Feng et al., [150] proposed a Factored Adaptation algorithm for non-stationary RL (FANS-RL), based on the factored non-stationary MDP model. FANS-RL aimed at learning a factored representation that encoded temporal changes of dynamics and rewards, including continuous and discrete changes. The learned representation then could be integrated with SAC [70] for identifying the optimal policy.

To improve generalization across environments of RL algorithms, Zhang et al., [177] focused on learning the state abstraction or representation from the varying observations. Based on block MDPs shown in Fig. 13, where sets of environments with a shared latent state space and dynamics structure over that latent space, they proposed a method using invariant causal prediction to learn the model-irrelevance state abstractions, which enable to generalize across environments with a shared causal structure. The model-irrelevance state abstraction in Fig. 13 must include both variables x^1, x^2. Based on the sparsity property in the dynamics that each next state variable only depends on a small subset of current state and action variables, Tomar et al., [178] proposed a model-invariant state abstraction learning method. Within the MBRL setting, their method allowed for the generalization

to unseen states. To further learn invariant representations for effective downstream control, Zhang et al., [146] proposed a method called Deep Bisimulation for Control (DBC) to learn robust latent representations which encoded task-relevant information, using bisimulation metrics without reconstruction. Intuitively, their focus was on the components in the state space that causally affected current and future rewards. They also proved the value bounds between the optimal value function of the true MDP and the one of the constructed MDP with the learned representations. Though Zhang et al., [146] learned the latent representations encoding the task-relevant information, they might elide the information which was irrelevant to the current task but relevant to another task, and they assumed the dynamics model between state abstractions was dense. To handle these issues, Wang et al., [152] proposed a method of learning causal dynamics for task-independent state abstraction. Specifically, they first decomposed state variables into three types: controllable variables, action-relevant variables and action-irrelevant variables. They then conducted conditional independence tests based on the conditional mutual information to identify the causal relationships between variables. By removing those unnecessary dependencies, they learned the dynamics that generalized well to unseen states, and finally derived the state abstraction. To generalize to unseen domains for the RL agents, Mozifian et al., [147] presented Intervention-Based Invariant Transfer learning algorithm (IBIT) in a multi-task setting, which regarded the domain randomization and data augmentation as forms of interventions to take away the irrelevant features from pixel observations. They derived an invariance objective via bisimulation matrics to learn the minimal causal representation of the environment which was relative to the relevant features of the solving task. Their method improved the zero-shot transfer performances of a robot arm in sim2sim as well as sim2real settings. To facilitate the agent to distinguish between causal relationships and spurious correlations in the environment of online RL, Lyle et al., [179] proposed a robust target-then-execute exploration algorithm to identify causal structures in MDP models, with theoretic motivations about the unbiased estimates of the value of any state-action pair.

Following the idea of meta-learning, Dasgupta et al., [180] used the model-free RL algorithm (A3C) to derive a method which enabled causal reasoning. They trained the agents to implement causal reasoning in three settings: observational, interventional and counterfactual, to gain rewards. This was the first algorithm that took a fusion between causal reasoning and meta model-free reinforcement learning.

In many scenarios like the control of a two-armed robot or a game of billiards, there may exist some sparse interactions where dynamics can be decomposed to locally independent causal mechanisms, as demonstrated in Fig. 14. Based on this finding, Pitis et al., [148] first introduced local causal models, which could be obtained by conditioning a subset of the state space from the global model. They then proposed an attention-based approach to discover such models for Counterfactual Data Augmentation (CoDA). Their method finally improved the sample efficiency for policy learning in batch-constrained and goal-conditioned settings. Different from [148] which used
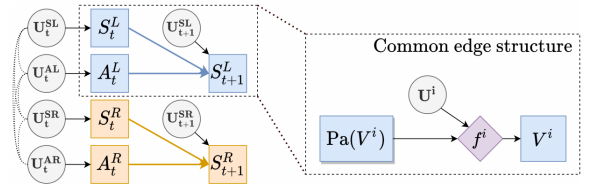

Fig. 14. An example of causal graphical illustration of MDP, where these state and action spaces were decomposed into two locally independent subspaces: $\mathcal{S} = \mathcal{S}^L \oplus \mathcal{S}^R, and \mathcal{A} = \mathcal{A}^L \oplus \mathcal{A}^R$ [148].

local causal models to augment the data counterfactually for the policy training, Seitzer et al., [149] used local causal models to detect situation-dependent causal influence, and to improve exploration and off-policy learning. In accordance with the intuition that only when the robot contacted the object, could the object be moved, they derived a measure of such local Causal Action Influences (CAI) with conditional mutual information. Integrating this measure into RL algorithms, their proposed CAI acted as an exploration bonus to perform active action exploration and to prioritize in experience replay for off-policy training.

In MBRL, there exist some endeavors that zero in on building the partial models with only part of the observatiosns. Considering the partial models, Rezende et al., [181] pointed out that previous works may not be causally correct, since part of the modeled observations may be confounded by other unmodeled observations. Hence, with MDP, they tried to figure out the information of states which affected the policy but belonged to unmodeled observations, and viewed such information as confounders. They proposed to learn a family of modified partial models for RL.

Transfer learning lays the groundwork for RL to deal with the data inefficiency and model-shift issues, but it may still suffer from the interpretability problem. Sun et al., [182] approached such a problem and presented a model-based RL algorithm which achieved explainable and transferable learning via causal graphical representations. The causal model aided in encoding the invariant and changing modules across domains. Specifically, based on the augmented source domain data, they firstly used a feasible causal discovery method, CD-NOD [34], to discover the causal relationships between variables of interest. Then, they leveraged the learned augmented DAG to infer the target model by training an auxiliary-variable neural network. They eventually transferred the learned model to train a target policy employing DQN or DDPG. To handle zero-shot multi-task transfer learning, Kansky et al., [183] introduced Schema Networks that are generative causal models for object-oriented reinforcement learning and planning. These networks allowed to explore multiple causes of events and reason backward through causes to achieve goals, with training efficiency and robust generalization.

2) POMDP: To bridge the gap between RL and causality, Gasse et al., [184] cast the model-based RL as a causal inference model. Specifically, they made the best of both online interventional and offline observational data on the basis of POMDP, and they proposed a generic method, aimed at uncovering a latent causal transition model, so as to infer
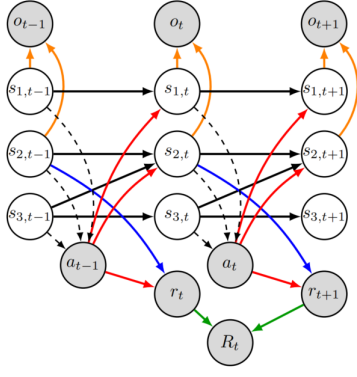
Fig. 15. An example of causal graphical illustration of POMDP, where grey nodes are observed variables while white nodes are unobserved ones [155].

the POMDP transition model via deconfounding. They also gave the correctness and efficiency proofs in the asymptotic case. Lu et al., [154] considered the RL problems which utilized the observational historical data to estimate policies when confounders might exist. They named the formalization as "deconfounded reinforcement learning" and also extended existing actor-critic algorithms to their deconfounding variants. Roughly speaking, they first identified a causal model from the purely observational data, discovering the latent confounders as well as estimating their causal effects on the action and reward. They thereafter deconfounded the confounders with the causal knowledge. Based on the learned and deconfounded model, they optimized the final policy. This was the first endeavour to combine confounders cases and the full RL algorithms.

Considering that perceived high-dimensional signals might contain some irrelevant or noisy information for decision-making problems in real-world scenarios, Huang et al., [155] introduced a minimal sufficient set of state representations (termed Action-Sufficient Representations, ASRs for short) that captured sufficient and necessary information for downstream policy learning. In particular, they first established a generative environment model which characterized the structural relationships between variables in the RL system, as illustrated in Fig. 15. And by constraining such relationships and maximizing the cumulative reward, they presented a structured sequential Variational Auto-Encoder method to estimate the environment model as well as the ASRs. This learned model and ASRs then would accelerate the policy learning procedure effectively.

Following the similar construction of the environment model in Fig. 15, Huang et al., [156] proposed a unified framework called AdaRL for the RL system to make reliable, efficient and interpretable adaptions to new environments. Specifically, they made the best of their proposed graphical representations to characterize what and where the changes adapt from the source domains to the target domain. That is, via learning factored representations that encoded changes in observation, dynamics and reward functions, they were capable of learning an optimal policy from the source domains. Thereafter, such a policy was adapted in the target domain with the identi-

fied changes. Experimental results demonstrated their superior policy adaptation performances with graphical representations. Sonar et al., [158] deployed an invariance principle to propose a reinforcement learning algorithm termed as the invariant policy optimization, with generalization ability beyond training domains. Following the POMDP setting in each domain, they learned a representation Φ mapping observations $o_t \in \mathcal{O}$ to hidden variables $h_t \in \mathcal{H}$ such that there existed a policy π mapping h_t to $a_t \in \mathcal{A}$ which was simultaneously optimal across domains. They finally obtained the invariant policy during training, which enabled generalization since such a policy found the causes of successful actions intuitively.

Aiming at the problems of data-efficiency and interpretability, Liang et al., [185], [186] presented an algorithm for POMDPs to discover time-delayed causal relations between events from a robot at regular or arbitrary times, and to learn transition models. They introduced hidden variables to stand for past events in memory units and to reduce stochasticity for observations. In such a way, they achieved training a neural network to predict rewards and observations. while identifying a graphical model between observations at different time steps. They then used planning methods for the model-based RL. Motivated by the fact that animals could reason about their behaviors through interactions with the environment, and discern the causal mechanism for the variations, Sontakke et al., [153] first proposed a causal POMDP model and introduced an intrinsic reward, called causal curiosity. This causal curiosity enabled the agent to identify the causal factors in the dynamics of the environment, giving interpretation about the learned behaviors and improving sample efficiency in transfer learning. To enable efficient policy learning in the environments with high-dimensional observations spaces, Zhang et al., [157] proposed a principled approach to estimate the causal states of the environment, which coarsely represented the history of actions and observations in POMDPs. Such causal states formed a discrete causal graph and helped predict the next observation given the history, which further facilitated the policy learning.

3) Bandits: As for the Multi-Armed Bandit (MAB) problem in the causal discovery setting, Lu et al., [159] proposed a causal bandit algorithm, namely Central Node UCB (CN-UCB), without knowing the causal structure. Their method was suitable for causal trees, causal forests, proper interval graphs, etc. Particularly, they first took an undirected tree skeleton structure as input, and output the direct cause of the reward. Then they applied the UCB algorithm on the reduced action set to identify the best arm. They finally showed theoretically that under some mild conditions, their method's regret scaled logarithmically with the number of variables in the causal graph. As stated, their method was the first causal bandit approach with better regret guarantee than standard MAB methods, in cases where the causal graph was not known in prior.

As for the contextual bandit problem, Saengkyongam et al., [160] solved the environmental shift issue in the offline setting, via the lens of causality. Their proposed multi-environment contextual bandit algorithms allowed for changes in the underlying mechanisms. They also delineated that in
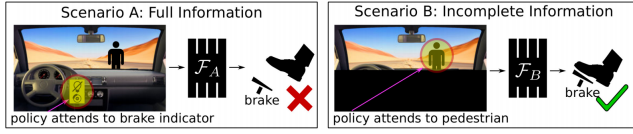
Fig. 16. Causal misidentification phenomenon: access to more information may yield worse imitation learning performance [163].

Fig. 17. An example of casual graphical illustration of IL, where expert demonstrations include observations x_t and actions a_t. States s_t consist of causal parents of the actions while η_t are noise representations that encapsulate any spurious correlations with the actions [166].

the case where unobserved confounders were present, the policy invariance they introduced could foster the generalization across environments under some assumptions. The heart of the invariance policy lied at the identification of the invariance set, which could be implied by the structure learning or be tested by the statistical method under distributional shifts [187]. Tennenholtz et al., [161] studied the problem of linear contextual bandits with observed offline data. They showed that the partially observed confounded information, which was characterized as linear constraints, could be estimated and utilized to online learning. Their proposed method could achieve the better overall regret.

4) IL: Taking the car driving as an example, De et al., [163] demonstrated that being oblivious to causality in the training procedure of IL would be damaging in learning policies, leading to a problematic phenomenon called "causal misidentification", as depicted in Fig. 16. Such a phenomenon implied that: access to more information can yield worse performance. To solve it, they proposed a causally motivated method. Firstly, they learned a mapping function from causal graphs to policies. Then through targeted interventions, either environment interaction or expert queries, they determined the correct causal models as well as policies. However, this algorithm [163] had high complexity especially in autonomous driving. Samsami et al., [60] hence proposed an efficient Causal Imitative Model (CIM) to handle inertia and collision problems from autonomous driving. Such undesirable problems were resulted from ignoring the causal structure of expert demonstrations. CIM first identified causes from the latent learned representation, and using Granger causality estimated the nest position. On the basis of the causes, CIM learned the final policy for the driving car. Wen et al., [168]–[171] considered a more specific causal confusion phenomenon, "copycat problem": the imitator tends to simply copy and repeat the previous expert action at the next time step. To combat this problem, they [168] proposed an adversarial IL approach to learn a feature representation that ignores the nuisance correlate from the previous action, still remaining the useful information for the next action; Wen et al., [170] gave up-weights to those keyframes which correspond to the expert action changepoints in the IL objective function, to learn copycat shortcut policies; Wen et al., proposed PrimeNet algorithm to improve rbustness and avoid shortcut solution. Since it was still an unresolved problem to well scale existing autonomous driving methods to real-world scenarios, Codevilla et al., [165] explored the limitations of behavior cloning and provided a new benchmark for investigations. For instance, they reported the generalization issue was partly due to the lack of a causal model. Data bias may lead to the inertia
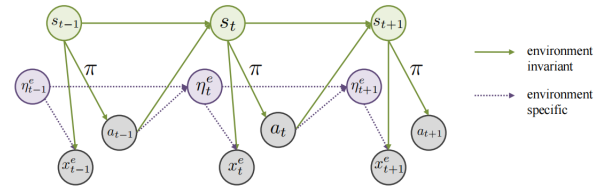
problem, suffering from causal confusion. Explicitly learning and employing a causal model is desirable in autonomous driving. Volodin et al., [188] and Lee et al., [189] learned causal structures between the states, and actions via interventions in the environment, enabling the policy learning procedure to take only cause states as inputs and preventing the spurious correlations of variables.

As for the generalization problems, Lu et al., [190] took a unified way to give the formalization, calling them the representation generalization, policy generalization and dynamics generalization. Following the behavior cloning approach to learn an imitation policy, they proposed a general framework to handle these generalization problems, with milder assumptions and theoretical guarantees. They identified the direct causes of a target variable and used such causes to predict an invariant representation which enabled generalizations. Bica et al., [166] aimed at learning generalizable imitation policies in the strictly batch setting from multiple environments, deploying to an unseen environment. To tackle the problems of spurious correlations between variables from expert's demonstrations, and of transition dynamics mismatch from multiple different environments, they learned a shared invariant representation, i.e., a latent structure that encoded the causes of the expert action, as well as a noise representation η which was environmentally specific, as demonstrated in Fig. 17.

As for the explanation problems, Bica et al., [164] bridged the gap between counterfactual reasoning and offline batch inverse RL, and proposed to learn the reward function with regard to the tradeoffs of the expert's actions. Their approach parameterized the reward functions according to the *what if* scheme, producing interpretable descriptions of sequential decision-making. Since they estimated the casual effects of different actions, counterfactuals could handle the off-policy nature of policy evaluation in the batch setting. As for IL methods with unknown causal structures, a general-purpose cognitive robotic IL framework centered on the cause-effect reasoning was first proposed by Katz et al. [191] in 2016. This framework focused on inferring a hierarchical representation of a demonstrator's intentions with causal reasoning methods, offering an explanation about the demonstrator's actions. In such a causal interpretation, top-level intentions, represented as the abstraction of skills, then could be reused to fulfill a specific plan in new situations [167], [191]. Further, during the planning procedure, Katz et al., [162] established causal plan graphs (CPGs) which modeled the causal relationships between hidden intentions, actions and goals. CPGs automat-

TABLE III
OVERVIEW CHARACTERISTICS OF EXISTING CRL METHODS.

Problems	CRL with Unknown Causal Information	CRL with Known Causal Information	Classical RL
Environment Modeling	compact causal graphs, confounding effects	compact causal graphs, confounding effects	fully connected graphs
Off-Policy Learning and Evaluation	action influences detection	confounding effects	no confounding effects
Data Augmentation	structural causal model learning	intervention and counterfactual reasoning	model-based RL
Generalization	causal discovery and causal invariance	invariance with causal information	invariance
Theory Analysis	causal identifiability and convergence[7]	convergence with causal information	convergence

ically gave causally-driven explanations of planned actions.

C. Summary

The sketch map of CRL framework is shown in Fig. 18, which gives an overview of possible algorithmic connections between planning and causality-inspired learning procedures. Causality-inspired learning could take place at three locations: in learning causal representations or abstractions (arrow a), learning a dynamics causal model (arrow b), and in learning a policy or value function (arrows e and f). Most CRL algorithms implement only a subset of the possible connections with causality, enjoying potential benefits in data efficiency, interpretability, robustness or generalization of the model or policy.
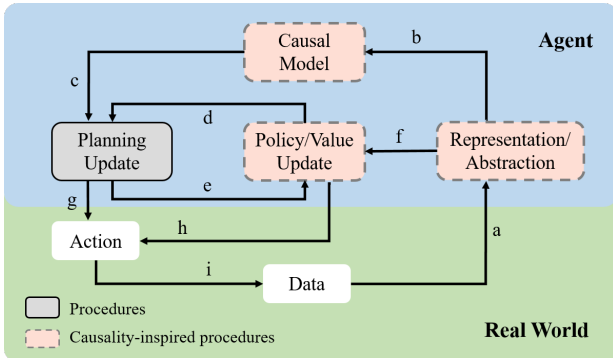

Fig. 18. A sketch map of CRL framework, that illustrates how causality information inspires current RL algorithms. This framework contains possible algorithmic connections between planning and causality-inspired learning procedures. Explanations of each arrow are: a) input training data for the causal representation or abstraction learning; b) input representations, abstractions or training data from real world for the causal model; c) plan over a learned or given causal model, d) use information from a policy or value network to improve the planning procedure, e) use the result from planning as training targets for a policy or value, g) output an action in the real world from the planning, h) output an action in the real world from the policy/value function, f) input causal representations, abstractions or training data from real world for the policy or value update.

According to different types of problems related to causality, we give an overview characteristics for CRL and traditional RL algorithms, with brief descriptions illustrated in Tab.III-C. In particular, as for the environment modeling, CRL usually employs the structural equation model to represent relationships between variables of interest as a compact causal graph, and it may consider the latent confounders; while classical RL usually treats the environment model as a fully directed graph, e.g., all states at time t would impact all states at time $(t + 1)$. As for the off-policy learning and evaluation, CRL with unknown causal information may evaluate the influences

of different actions, while that with known causal information usually investigates the confounding effects on the policy, by sensitivity analysis. Traditional RL would not model the confounding effects. As for the data augmentation problem, classical RL sometimes is based on the model-based RL, while CRL on the structural causal model. After learning such a model, CRL could perform counterfactual reasoning to achieve data augmentation. As for the generalization, classical RL attempts to explore invariance while CRL tries to leverage causal information to produce causal invariance, e.g., structural invariance, model invariance, etc. As for the theoretical analysis, classical RL often concerns convergence, including the sample complexity and regret bounds of the learned policy, or the model errors; CRL also concerns convergence but with causal information, and it focuses on causal structure identifiability analysis.

IV. EVALUATION METRICS

In this section, we list some commonly used evaluation metrics in causal reinforcement learning, which essentially originate from fields of RL and causality.

A. Policy

To evaluate the algorithm's performances, one often assesses the estimated policy via the average cumulative reward, average cumulative regret, fractional success rate, probability of choosing the optimal action, or their variants. Cumulative rewards are averaged over episodes (or over runs) with the defined reward function in the training or testing phases [90], [130], [150], [154], [180], [182], [186]. Cumulative regrets, i.e., the reward difference between the optimal policy and the agent's policy, are computed and averaged over episodes (or over runs) [12], [110], [111], [115], [131], [161]. The Fractional Success Rate (FSR) is usually deployed in evaluation in robotic manipulation or related tasks, e.g., picking up an object to a target location. FSR measures the overlapping success ratio between the object and the target [87], [192]. In causal bandit problems, the probability of choosing the optimal arm can be used as the evaluation measurement [12]. It measures the percentage of the optimal action over all time steps in each episode [154].

B. Model

In causal reinforcement learning with the environment model, one enables to evaluate the quality of the learned dynamics. Such a quality is measured by the transition loss of
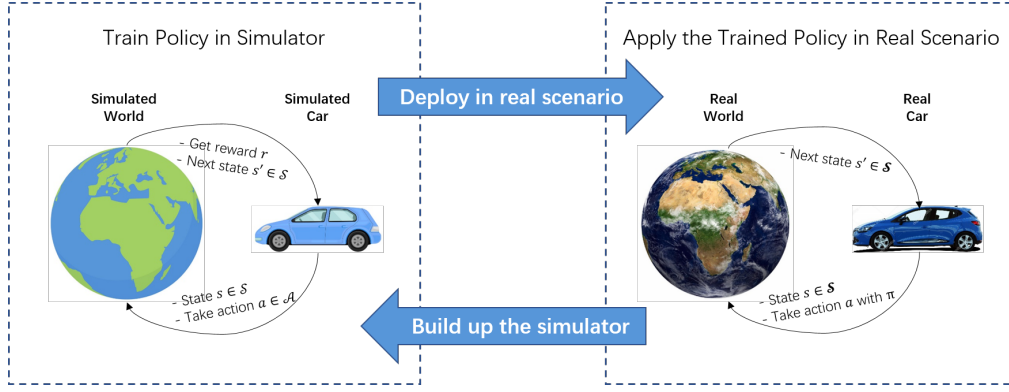
Fig. 19. The overall framework of real world applications from CRL methods. On the one hand, a simulator is built up or corrected from the real scenario; while on the other hand, one may deploy the trained policy from simulator into real scenario.

the dynamics, between the ground truth states and predicted ones [89], [152], [177], e.g.,

$$Errors_M = \sum_{episodes} \sum_t (s_t - \hat{s}_t)^2,$$

where s_t stand for ground truth states while $\hat{s}_t$ are the predicted ones from the dynamic model. Normally, a low transition loss helps learn a good policy.

C. Causal Structure

When identifying causal structures in causal reinforcement learning, one may evaluate the accuracy of the structure learning. As such, existing works often compute the edge accuracy (formally called precision) with the ground truth graph [143], [152]. In causal discovery domain, recall, and F1 values are employed for the evaluation as well. They are defined as below,

$$Precision = \frac{TP}{TP + FP},$$
$$Recall = \frac{TP}{TP + FN},$$
$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

where TP, FP, FN are true positives, false positives, and false negatives, respectively. TP correctly indicates the presence of an edge in the graph; FP wrongly indicates that an edge is present while FN wrongly indicates that an edge is absent in the graph.

V. APPLICATIONS

Real world applications avoid errors from risky exploration and exploitation, while on the contrary CRL methods attain the trial-and-error mechanism. To break up the contradicts between CRL and a specific real-world application, a credible strategy is to build up an authentic simulator based on some real data and domain knowledge about the model dynamics, then design the objectives for the agent and train the policy network in the simulator, and eventually deploy the trained policy in the real world with further improvement. The overall framework of real world applications from CRL methods is

depicted in Fig.19. Despite its fantastic and potential future, CRL still lies in infancy. Most current CRL methods basically study the performances in the simulator, with few of them in real-world scenarios. In this section, we review applications of the CRL methods in various fields, including but not limited by robotics, games, medical treatment, and autonomous driving, recommendation, etc., most of which are evaluated in simulators.

A. Robotics

Robotics has been applied flourishing in reinforcement learning [7], [193], [194]. And researchers has attempted to apply CRL as well in robotics to verify the performances of methods, since CRL can be incorporated in a straightforward way for robots to conduct some tasks via acting intellectually in complex open-world environments.

Some prior works have explored causal IL robotic framework for learning and generalizing robotic skills [191]. These methods infer a hierarchical representation of a demonstrator's intentions, which is used to the planning procedure [162], [167], [191]. [172] takes the best of the learned causal structure to foster policy learning for a given task of service robotics. [146] evaluates their causal dynamics and policy learning in the *robosuite* simulation framework. [189] focuses on robot manipulation tasks of block stacking and crate opening for policy transfer verification. [150] proposes a factored adaptation algorithm for non-stationary RL, encoding the temporal changes of dynamics and rewards, and considers two robotic Sawyer manipulation tasks (Reaching and Peg) and a minitaur robot that learns to move at a target speed. [185], [186] discover time-delayed causal relations between events from a robot at regular or arbitrary times, while learning the optimal policy in the model-based RL framework. Methods are evaluated in two robotic experiments in the *Gazebo* simulator, and a real robot for the painting task. [158] proposes an invariant policy optimization algorithm and utilizes the colored-key problem for evaluation where the robot has to learn to navigate to the key, use it to open the door, and then navigate to the goal. They further consider a challenging task where the robot learns to open a door with different door's phisical parameters.
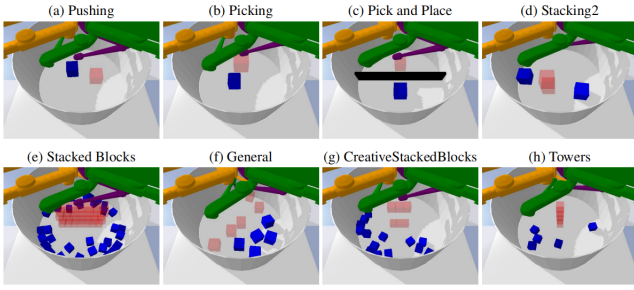
Fig. 20. Some example tasks provided in the benchmark of *CausalWorld* where the opaque red blocks stand for the goal shape while the blue ones are available blocks to be processed [192].

Other relative works with robotic manipulation applications include [147], [148]

Other prior works have explored robotic platforms for constructing a family of robotic manipulation environments, i.e., *CausalWorld* [192]. It is an open-source platform which allows for causal structure learning and transfer learning on new robotic environments. The robot in *CausalWorld* is a 3-finger gripper, and each finger has three joints. The robot learns to move objects to specified target locations. Fig. 20 depicts some example tasks provided in the benchmark where the opaque red blocks stand for the goal shape while the blue ones are available blocks to be processed. [87], [153] use *CausalWorld* to verify their methods' transfer learning and zero-shot generalizability in down-stream tasks.

For multi-joint dynamics control, one of the most widely applicable robotic platforms in RL belongs to MuJoCo. MuJoCo stands for Multi-Joint dynamics with Contact, which is a general-purpose physics engine consisting of 10 simulated environments of continuous controlling [195]. Researchers validate their approaches with improved performances of policy learning on MuJoCo-based Humanoid [51], [178], Hopper [163], Inverted Pendulum [143], Half Cheetah [51], [126], [177], Ant [126], Walker [51], etc.

B. Games

Games enable to provide excellent testbeds for RL including CRL agents, since games, especially video games can vary enormously but still present challenging and interesting objectives for humans.

As for Atari games, [183] introduces Schema Networks to explore multiple causes of events for reasoning, resulting in high efficiency and zero-shot transfer on variations of game Breakout. Variations of such a game include standard version, versions with a middle wall, offset paddle, etc. [156] evaluates their proposed AdaRL algorithm that adapts to changes across domains, in the modified Atari Pong environments. They set changes in the orientations, colors, or reward functions in Pong to test robustness to change factors. [163] uses Atari games to state causal misidentification phenomenon indeed adversely influence imitation learning performance. Works that utilize various Atari games to evaluate the performances, sample efficiency or generalization ability include [89], [148], [157], etc. As for Sequential Social Dilemmas (SSDs) which are

spatially and temporally extended multi-agent games, [103] chooses a public goods game *Cleanup*, and a tragedy-of-the-commons game *Harvest* to empirically demonstrate the enhancement of the proposed social influence reward. As for the StarCraft games, [102] applies decentralized StarCraft micromanagement to evaluate their COMA method's efficacy, while [132] adopts Starcraft II to evaluate the faithfulness of their proposed action influence models in explainability of behaviors. Peng et al., [151] employ 2d-Minecraft [174] and simplified sandbox survival games Eden [175] to verify the high quality of subgoal hierarchy in HRL driven by causality. 2d-Minecraft and Eden are two types of games with complex environments and especially with sparse rewards. In addition, VizDoom, grid-world, and Box2D games have also been widely investigated in CRL for performance verification [89], [126], [128], [133], [155], [157], [188], etc.

C. Healthcare

Applying reinforcement learning related techniques in healthcare has been appealing but it also poses some significant challenges since it is concerned about one's safety and health [194], [196]. Performing online exploration on patients is largely impeded while some medical treatments might be costly, dangerous, or even unethical. Further, often only a few records of each patient for the healthcare data are available, and patients may give different responses to the same treatments.

Regarding to heterogeneity across individuals, [88] proposes to model the personalized treatment and general treatment, and investigates the method's performances on the real-world Medical Information Mart for Intensive Care-III (MIMIC-III) database with Sepsis-3. MIMIC-III is an influential dataset in healthcare treatment, which contains 6631 medical records from Intensive Care Units (ICUs) [197]. This dataset has also been used by [164], [166]. Outside of MIMIC-III dataset, [121] considers a dataset with a total of 183, 487 mothers who delivered exactly two births during 1995 and 2009 in the Commonwealth of Pennsylvania. Their aim is to estimate the DTR from retrospective observational data using a time-varying instrumental variable. [118] focuses on a case study on the parallel Women's Health Initiative (WHI) clinical data to illustrate improvement of their confounded method in policy learning. [117] studies the International Stroke Trial (IST), demonstrating their safety guarantees towards confounded selection. [62], [119] test the model of multi-stage treatment regimes for lung cancer or dyspnoea. Regarding to the safety and sample scarcity issues, other prior works construct a synthetic medical environment for performances investigation [98], [99], [137]. [97] applies a simple Sepsis simulator to construct a medical environment for illustration, while [94] uses simulators of Sepsis and autism management to certify the robustness of unobserved confounding. [93] performs a mobile health study where the synthetic dataset is generated by a generative model and from a mobile health application that monitors the blood glucose level of type 1 diabetic patients. [125] constructs an assistive-healthcare environment for the simulated robots to offer physical assistance to disabled persons.

D. Autonomous Driving

Autonomous driving is a particular suitable application in CRL, since safety and efficiency are two major barriers, and RL along with causality knowledge can be regarded as a potentially promising tool for enabling safe, effective learning in autonomous driving. As the widespread availability of high-quality demonstration data, imitation learning is a popular approach in autonomous driving [196].

Through the real-world driving setting, Codevilla et al. [165] and de Haan et al. [163] demonstrate the problems of causal misidentification and causal confusion in IL that accessing history or spurious correlations yields better validation performance, but worse actual driving performance. They overcome the problems by learning the correct causal model in the environment to find the correct policy. To verify the causal transfer ability for IL under the sensor shift, Etesami and Geiger [123] employ a real-world dataset that contains recordings of human-driven cars by drones that fly over several highway sections in Germany. Zhang et al. [122] test their causal imitation method using HighD dataset, which contains drone recordings of human-driven cars as well.

Since most CRL has not yet been applied in actual real-world autonomous driving vehicles, simulated self-driving environments have been gaining increasing popularity, e.g., CARLA [60], [146], [165], Car Racing [155], Toy Car Driving [143], etc. CARLA environment consists of different towns with different lengths of accessible roads and number of cars in urban areas. Apart from the original benchmark, Codevilla et al. [165] devise a larger scale CARLA benchmark, called NoCrash, to further evaluate their method's efficacy in complex conditions with dynamic objects. Zhang et al. [146] construct a highway driving scenario for evaluation. Car Racing is a continuous task for a car with three continuous actions, and it has been used to validate the efficiency and efficacy of Action-Sufficient state Representations (ASRs) in improving policy learning [155]. Zhu et al. [143] take the best of Toy Car Driving environment, which is a typical RL environment for a car to carry out tasks of avoiding obstacles or navigating, to evaluate the performances of causal structure learning. State variables include the car's direction, the velocity, and the position, etc.

E. Recommendation

Recommendation systems deal with the problem of recommending multiple items in sequence to the user on the basis of the user's delayed feedback [198]. For instance, to pursue revenue maximization, online car hailing systems often recommend sequentially and adaptively different levels of riding coupons to passengers, according to their past car hailing behaviors and feedback. Due to the difficulty for the simulator to offer fully observed environment information of the real world and the existence of unobserved confounding variables, CRL solutions could be appealing.

In terms of the simulator of recommendation systems, Shi et al. [199] establish Virtual-Taobao for better commodity search in Taobao, which is one of the largest online retail platforms. Their simulator allows to train a recommendation policy with no physical sampling costs. In terms of the CRL approaches in recommendation systems, Shang et al. [200] propose a deconfounded multi-agent environment reconstruction (DEMER) method to achieve environment reconstruction with latent confounders. They apply an artificial application of driver program recommendation and a real application of Didi to verify the efficacy of DEMER. Lu et al. [110] use the email campaign data from Adobe, where the reward is obtained after the commercial links inside the email are clicked by the recipient. They demonstrate that their proposed bandit methods with causal information yield lower average regrets than those pure UCB or TS methods. Li et al. [111] adopt Weixin A/B testing data with associated logs, as well as Yahoo's news recommendation data, to show the superiority of using logged data and online feedbacks when making decisions. Tennenholtz et al. [125] conduct experiments to test the severity of covariate shifts in the confounded imitation setting on the RecSim Interest Evolution environment, which is a slate-based recommender system for simulating sequential user interaction. Shi et al. [93] present an off-policy value estimation algorithm in the confounded MDP setting, and they justify their method with a real-world dataset from a ride hailing company. This dataset is about a recommendation program applied to customers in regions where there are more taxi drivers than the call orders, with which they aim at evaluating the coupon recommendation policies.

F. Others

RL has many other applications though, we here list existing works on CRL in the domain of classical control. Such domains as financial economy, natural language process, etc. are in need of exploration in CRL.

In terms of classical control, Sun et al. [182] perform experiments on Pendulum, Mountain Car, and Cart Pole environments and simulated quadcopter recordings to verify the robustness of their framework. Zhang et al. [177] give an illustration of the benchmark Cartpole and show they enable to learn Model-Irrelevance State Abstractions (MISA) while training a policy. Chen et al. [51] modify environments of Mountain Car, Cart Pole, and Catch to demonstrate how IV variables could address the confounding bias. De Haan et al. [163] create a variant of Mountain Car confounded testbed to indicate the necessity of disentanglement of the state representation, while Shi et al. [105] consider the Cart Pole environment for evaluation.

VI. OPEN SOURCES

A. Codes

We summarize codes with prior causal information and unknown causal information in Tab. IV and Tab. V, respectively.

B. Tutorials, Reports or Surveys

One of the most informative and comprehensive tutorials belongs to the one presented by Bareinboim [15] at ICML 2020: https://crl.causalai.net/. He introduced the foundations

TABLE IV
PART OF CRL CODES WITH PRIOR CAUSAL INFORMATION

Methods	Sources
Causal TS [12]	https://github.com/ucla-csl/mabuc
IVOPE [85]	https://github.com/liyuan9988/IVOPEwithACME
DFPV [53]	https://github.com/liyuan9988/DeepFeatureProxyVariable/
COPE [93]	https://github.com/Mamba413/cope
off_policy_confounding [94]	https://github.com/StanfordAI4HI/off_policy_confounding
RIA [95]	https://github.com/CR-Gjx/RIA
Gumble-Max SCM [97]	https://github.com/clinicalml/gumbel-max-scm
Confounded-POMDP [105]	https://github.com/jiaweihhuang/Confounded-POMDP-Exp
Causal Bandit [109]	https://github.com/finnhacks42/causal_bandits
POMIS [113]	https://causalai.net/
Unc_CCB [115]	https://openreview.net/forum?id=F5Em8ASCosV
OptZ [116]	https://github.com/CausalML/LatentConfounderBalancing
CRLogit [117]	https://github.com/CausalML/confounding-robust-policy-improvement
CTS [125]	https://openreview.net/forum?id=w01vBAcewNX
ResidualIL [126]	https://github.com/gkswamy98/causal_il

TABLE V
PART OF CRL CODES WITH UNKNOWN CAUSAL INFORMATION

Methods	Sources
CIM [60]	https://github.com/vita-epfl/CIM
ICIN [129]	https://github.com/StanfordVL/causal_induction
causal InfoGAN [144]	https://github.com/thanard/causal-infogan
NIE [145]	https://gitlab.compute.dtu.dk/tuhe/causal_nie
DBC [146]	https://github.com/facebookresearch/deep_bisim4control
IBIT [147]	https://github.com/melfm/ibit
CoDA [148]	https://github.com/spitis/mrl
CAI [149]	https://github.com/martius-lab/cid-in-rl
CDL [152]	https://github.com/wangzizhao/CausalDynamicsLearning
Deconf. AC [154]	https://github.com/CausalRL
AdaRL [156]	https://github.com/Adaptive-RL/AdaRL-code
IPO [158]	https://github.com/irom-lab/Invariant-Policy-Optimization
linear CB [161]	https://github.com/guytenn/Bandits_with_Partially_Observable_Offline_Data
causal misidentification [163]	https://github.com/pimdh/causal-confusion.
CILRS [165]	https://github.com/felipecode/coiltraine/blob/master/docs/exploring_limitations.md
ICIL [166]	https://github.com/ioanabica/Invariant-Causal-Imitation-Learning
cause-effect IL [167]	https://github.com/garrettkatz/copct
fighting-copycat [168]	https://github.com/AlvinWen428/fighting-copycat-agents
keyframe [170]	https://tinyurl.com/imitation-keyframes
PrimeNet [171]	https://github.com/AlvinWen428/fighting-fire-with-fire
causal-rl [184]	https://github.com/causal-rl-anonymous/causal-rl

of causal inference and reinforcement learning as well as discussed six fundamental and prominent CRL tasks. Such tasks are i) generalized policy learning; ii) when and where to intervene; iii) counterfactual decision making; iv) generalizabililty and robustness of causal claims; v) learning casual models; and vi) casual imitation learning. Other technical reports, blogs or surveys are shown as below:

- Lu [16] was motivated by the current development in healthcare and medicine, and introduced causal reinforcement learning. He summarized the advantages of CRL as *data efficiency* and *minimal change*.
- A company called causalLens briefly summarized current research trend of CRL in cases where there are latent confounders or in deep model-based CRL: https://www.causalens.com/technical-report/causal-reinforcement-learning/.
- Ness shared his bet on CRL as the next killer application of marketing science within the next ten years: https://towardsdatascience.com/my-bet-on-causal-reinforcement-learning-d94fc9b37466.
- Grimbly et al., [17] surveyed on causal multi-agent reinforcement learning. Grimbly also discussed the six tasks from [15] with surveying relevant CRL algorithms: https://stjohngrimbly.com/causal-reinforcement-learning/.
- Bannon et al., [18] surveyed on causal effects' estimation and off-policy evaluation in batch reinforcement learning.
- Kaddour et al., [19] surveyed on causal machine learning that included a chapter of CRL, summarizing causal

bandits algorithms, etc.

C. Competitions

Apart from the fruitful competitions with reinforcement learning that CRL algorithms could be exploited, there exists another competition related to both causality and reinforcement learning. That is the Learning By Doing: Controlling a Dynamical System using Control Theory, Reinforcement Learning, or Causality. It is a competition of NeurIPS 2021, which aims at encouraging cross-disciplinary discussions among researchers from fields of control theory, RL and causality. Two tracks were designed. In Track CHEM, the goal is to find optimal policies for the concentrations of chemical compounds to ensure a specific target concentration on one of the products. In Track ROBO, the goal is to learn optimal policies for a robot arm, such that the tip of the robot follows a target process via sequentially interaction.

For detailed background information or solutions, please refer to: https://learningbydoingcompetition.github.io/ [201].

VII. FUTURE DIRECTIONS

We categorize current CRL approaches into two sides, according to whether their causal information is given or not. While it is not that easy to attain and take the best of causal information from pure observed data, in this section, we discuss some future directions in CRL from the following 4 aspects: i) knowledge-based: how to extract knowledge when causal information is unknown, ii) memory-based: how to store and remember causal information, iii) cognition-based: how to exploit prior causal information from related tasks or experts, and iv) application-based: how to apply CRL algorithms in real worlds.

A. Knowledge-based causal reinforcement learning

Knowledge-based CRL aims at incorporating knowledge into CRL, and often involves the procedure of extracting knowledge from data. When encountering new environments or tasks, updating knowledge is more than needed. Knowledge may have various forms (e.g., causal structure, causal ordering, causal effects, causal features, etc.) and may be incorporated in the model, value function, reward function, etc.

Taking the environment model for instance, how to ensure that the estimated environment model with knowledge is consistent with the actual real-world one becomes a core challenge (distributional shift problem) [8]. To handle this challenge, one possible solution is to consider latent confounders in CRL. Since there is a great possibility for the existence of latent confounders in many applications, studying confounders enables current CRL methods to learn a more accurate model or policy, which mitigates the shortcomings from the differences between the learned policy and the optimal policy. Though there exist some works on confounders in CRL that we review in Section III, it is suggested that developing general methods with weaker assumptions to detect or deploy latent confounders and eliminate their effects, may be potentially relevant for effective, robust, and interpretable causal reinforcement learning. Another possible solution to distributional shift is to take into account the non-stationarity of causal structures in the model, e.g., the causal relations between states, actions, and rewards. Such non-stationarity may be originated across multiple time steps, multiple domains, multiple modals of data, multiple tasks or even multiple agents, which implies that causal structures in different time steps, domains, modals, tasks, or agents may change. For instance, in robotic manipulation environments [192], the causal structure would change whenever a robotic arm touches or picks one of the objects; For the Pushing and Picking tasks, causal relations between actions and rewards are non-stationary; For the identical task in the same environment, different types of robotic arms (agents) may own different characteristics, rendering the same action by arms towards the same object possibly produce different

dynamics, rewards or policy generation mechanisms; There also may exist cooperative, defensive or adversarial agents in the multi-agent setting. However, the questions of how such non-stationarity of the knowledge affects the dynamics, rewards or policy functions, what real effects are and how to deal with them, remain a large region to be explored.

B. Memory-based causal reinforcement learning

Memory-based CRL expects to be possessed of powerful memory that preserves the sequences or the distributions of state, action as well as reward, so as to remember previously-visited regions (e.g., interesting states, actions, etc.) and eventually achieve more effective and efficient policy learning in an intelligent design. Current RL algorithms, based on recurrent neural networks to integrate information over long time periods, usually deploy attention or differentiable memory techniques to store and retrieve successful experiences for rapid learning [202], [203].

To spawn memory-based CRL methods, one may gain advantages from the combination of both on-policy and off-policy algorithms, while avoiding getting stuck in local optima and improving convergence rates. To keep memory and prevent forgetting those skills that ever trained, how to design an experience replay scheme to take the best of memory information for transfer learning, deserves exploration in CRL. Alternatively, the setting of lifelong CRL is also another interesting research direction to prevent forgetting based on memory replay.

C. Cognition-based causal reinforcement learning

Cognition-based CRL attempts to exploit previously acquired or newly updated cognition capabilities from related tasks or experts, to speed up learning in CRL with high performances. It follows the logic from the brain, and involves the processes of perception, understanding, judgment, evaluation, abstract thinking and reasoning, etc [204].

Intervention and counterfactual analysis for CRL are recent emerging paradigms that utilize the existing cognitive model to improve sample efficiency and facilitate learning. Regarding intervention issue, with causal models taken into account in the sequential decision-making setting, the interesting avenues for research include questions of when, where and how to intervene variables of interests to lead to optimal policies, based on perception with less assumptions. These ways could be used as a heuristic for a possible refinement of the policy space based on topological constraints [15], [114], [136], as well as for a renewal of the causal model by correcting the causal direction or causal knowledge. Regarding counterfactual analysis issue, since it involves thinking about possibilities or imagination in the cognitive process, it plays vital roles in investigating the outcomes from some dangerous, unethical or even impossible action spaces in safety critical applications. Causality offers tools for modeling such uncertain scenarios for analysis, thus enjoying sample efficiency and policy effectiveness.

A key to the success of cognition-based CRL belongs to the generalization capability of methods, including learning generalizable models as well as training generalizable policies

---

[8]Existing RL algorithms, to deal with the distributional shift problem, are mainly via policy constraints or uncertainty estimation [196].

towards other unseen testing environments, tasks or agents. Ahmed et al., [192] empirically find that in *CausalWorld*, an agent's generalization capabilities are related to the experience gathered at the training phase. Based on the training experience, how to characterize the causal structural invariances and the variances that are shared or specific across domains, tasks or agents from the acquired knowledge, is still a potentially fruitful research direction. In addition to the structural invariances, invariant causal features or state abstractions may also be effective alternatives to aid in generalizability. Relevant techniques are meta-learning, domain generalization, etc. Besides, causal models open room for hierarchical RL to adaptively form the high-level policy instead of manually dividing the goals, which is an interesting area of ongoing research for generalization capability.

D. Application-based causal reinforcement learning

Application-based CRL focuses on applying current CRL algorithms in real-world applications such that the key requirements and challenges in the industrial community could be tackled. Possible applications include games, robotics (e.g., service robots, medical robots, robotic arms, aerial vehicle), military, healthcare (e.g., medical treatments), financial trading, product recommendation, urban transportation (e.g., autonomous driving), smart cities (e.g., electricity allocation), etc. Taking autonomous driving for instance, a car is required to reach the goal from the start safely and reliably, obeying the traffic rules with no accidents and enjoying real-time performances. To facilitate real-world applications of CRL, causal information from the simulator (or environment dynamics) or datasets, whether from prior knowledge or not, lacks investigation, which can be further explored in future work.

Building hand-crafted simulators has been widely adopted in RL, such as autonomous driving [205], [206], robotic control [192], etc.. Simulators are beneficial to generate specific scenarios in real-world applications that are rare for robust policy training [73], e.g., attack-defense game or cooperative-adversarial scenarios in multi-agent settings, etc. However, since it is hard for simulators to obtain high fidelity compared with real-world scenarios, performances may be unreliable and sensitive to the simulator-reality gap when utilizing the trained policy network into real tasks. Thus, accompanied with the simulators, we have to look back to the principal overall functionality and demands from each of the applications to reduce as much simulator-reality gap as possible; accompanied with the algorithms, we could consider the techniques for generalization capability in cognitive-based CRL, to reduce the unfavorable effects from the gap and to improve the human-machine interaction capability.

While current CRL works place considerable value on design of simulators, new algorithms and theory, much practical progress in real-world applications might have been and will be driven by advances in datasets, similar to the fields of computer vision and natural language process. Collecting representative, large, and diverse datasets is often far more significant than utilizing the most advanced methods if factoring in real-world applications [196]. Whereas most standard RL methods belong to interactive learning with the environment, collecting large and diverse datasets is often impractical, costly, or unethical, due to safety or privacy concerns. The machine learning and CRL communities shed light on the facts that i) Applications could well benefit from the real-world, large and diverse datasets; ii) Methods integrated with offline data enable to well handle a wide range of real-world problems. Hence, careful collections of available datasets, as well as the development of data-driven or data-fusion CRL algorithms, could potentially induce a series of breakthroughs towards more efficient, effective and practical CRL techniques in the near future.

VIII. CONCLUSIONS

We have witnessed the amazing and flourishing progress of causal reinforcement learning in recent years. In this survey, we take a comprehensive review on causal reinforcement learning, where existing CRL methods roughly fall into two categories. One category is based on the unknown causality information which has to learn before training policies, while the other one relies on the known causality information. For each category, we analyze the reviewed methods according to their model settings, i.e., settings of MDP, POMDP, Bandits, DTR or IL. We further offer evaluation metrics, open sources and some occurring applications in CRL. We eventually summarize future directions from the following four aspects: i) knowledge-based; ii) memory-based; iii) cognition-based; and iv) application-based CRL, which may empower CRL with a new era of advances.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[5] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman *et al.*, "Human-level performance in 3d multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859–865, 2019.

[6] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[7] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[8] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *International Conference on Machine Learning*. PMLR, 2016, pp. 49–58.

[9] Y. Lin, Y. Liu, F. Lin, P. Wu, W. Zeng, and C. Miao, "A survey on reinforcement learning for recommender systems," *arXiv preprint arXiv:2109.10665*, 2021.

[10] Y. Sun, F. Zhuang, H. Zhu, Q. He, and H. Xiong, "Cost-effective and interpretable job skill recommendation with deep reinforcement learning," in *Proceedings of the Web Conference 2021*, 2021, pp. 3827–3838.

[11] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–36, 2021.

[12] E. Bareinboim, A. Forney, and J. Pearl, "Bandits with unobserved confounders: A causal approach," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[13] P. Jonas, J. Dominik, and S. Bernhard, *Elements of Causal Inference: foundations and learning algorithms*. MIT Press, 2017.

[14] J. Pearl, *Causality: Models, Reasoning, and Inference*. New York, NY, USA: Cambridge University Press, 2000.

[15] E. Bareinboim, "Causal reinforcement learning," *Tutorials at the Thirty-ninth International Conference on Machine Learning*, 2020.

[16] C. Lu, "Introduction to causal reinforcement learning," *Blogpost at causallu.com*, 2018.

[17] S. J. Grimbly, J. Shock, and A. Pretorius, "Causal multi-agent reinforcement learning: Review and open problems," in *Cooperative AI Workshop in Advances in Neural Information Processing Systems*, 2021.

[18] J. Bannon, B. Windsor, W. Song, and T. Li, "Causality and batch reinforcement learning: Complementary approaches to planning in unknown domains," *arXiv preprint arXiv:2006.02579*, 2020.

[19] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, "Causal machine learning: A survey and open problems," *arXiv preprint arXiv:2206.15475*, 2022.

[20] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[21] J. S. Sekhon, "The neyman-rubin model of causal inference and estimation via matching methods," *The Oxford handbook of political methodology*, vol. 2, pp. 1–32, 2008.

[22] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, vol. 10, p. 524, 2019.

[23] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2011, pp. 804–813.

[24] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.

[25] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT press, 2000.

[26] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.

[27] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour, "Generalized score functions for causal discovery," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1551–1560.

[28] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol. 7, no. Oct, pp. 2003–2030, 2006.

[29] P. L. Spirtes, C. Meek, and T. S. Richardson, "Causal inference in the presence of latent variables and selection bias," *arXiv preprint arXiv:1302.4983*, 2013.

[30] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social science computer review*, vol. 9, no. 1, pp. 62–72, 1991.

[31] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 507–554, 2002.

[32] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine learning*, vol. 20, no. 3, pp. 197–243, 1995.

[33] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[34] B. Huang, K. Zhang, J. Zhang, J. D. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, "Causal discovery from heterogeneous/nonstationary data." *J. Mach. Learn. Res.*, vol. 21, no. 89, pp. 1–53, 2020.

[35] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," *Advances in Neural Information Processing Systems*, vol. 21, 2008.

[36] K. Zhang and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2009, pp. 647–655.

[37] D. Entner and P. O. Hoyer, "On causal discovery from time series data using fci," *Probabilistic graphical models*, pp. 121–128, 2010.

[38] D. Malinsky and P. Spirtes, "Causal structure learning from multivariate time series in settings with unmeasured confounding," in *Proceedings of 2018 ACM SIGKDD workshop on causal discovery*. PMLR, 2018, pp. 23–47.

[39] A. Hyvärinen, S. Shimizu, and P. O. Hoyer, "Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 424–431.

[40] A. Gerhardus and J. Runge, "High-recall causal discovery for auto-correlated time series with latent confounders," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 615–12 625, 2020.

[41] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[42] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[43] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[44] M. Caliendo and S. Kopeinig, "Some practical guidance for the implementation of propensity score matching," *Journal of economic surveys*, vol. 22, no. 1, pp. 31–72, 2008.

[45] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[46] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International Conference on Machine Learning*. PMLR, 2016, pp. 3020–3029.

[47] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the national academy of sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.

[48] X. Nie and S. Wager, "Quasi-oracle estimation of heterogeneous treatment effects," *Biometrika*, vol. 108, no. 2, pp. 299–319, 2021.

[49] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein, "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models," in *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000, pp. 1–94.

[50] Z. Tan, "A distributional approach for causal inference using propensity scores," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1619–1637, 2006.

[51] Y. Chen, L. Xu, C. Gulcehre, T. L. Paine, A. Gretton, N. de Freitas, and A. Doucet, "On instrumental variable regression for deep offline policy evaluation," *arXiv preprint arXiv:2105.10148*, 2021.

[52] J. Li, Y. Luo, and X. Zhang, "Causal reinforcement learning: An instrumental variable approach," *Available at SSRN 3792824*, 2021.

[53] L. Xu, H. Kanagawa, and A. Gretton, "Deep proxy causal learning and its application to confounded bandit policy evaluation," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 26 264–26 275.

[54] W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen, "Identifying causal effects with proxy variables of an unmeasured confounder," *Biometrika*, vol. 105, no. 4, pp. 987–993, 2018.

[55] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998.

[56] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[57] R. Nian, J. Liu, and B. Huang, "A review on reinforcement learning: Introduction and applications in industrial process control," *Computers & Chemical Engineering*, vol. 139, p. 106886, 2020.

[58] E. L. Thorndike, "Animal intelligence," *Nature*, vol. 58, no. 1504, pp. 390–390, 1898.

[59] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," *Advances in Neural Information Processing Systems*, vol. 20, no. 1, pp. 96–1, 2007.

[60] M. R. Samsami, M. Bahari, S. Salehkaleybar, and A. Alahi, "Causal imitative model for autonomous driving," *arXiv preprint arXiv:2112.03908*, 2021.

[61] S. A. Murphy, "Optimal dynamic treatment regimes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 2, pp. 331–355, 2003.

[62] J. Zhang and E. Bareinboim, "Designing optimal dynamic treatment regimes: A causal reinforcement learning approach," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 012–11 022.

[63] T. M. Moerland, J. Broekens, and C. M. Jonker, "Model-based reinforcement learning: A survey," *arXiv preprint arXiv:2006.16712*, 2020.

[64] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1889–1897.

[65] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[66] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[67] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1995–2003.

[68] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," *Advances in Neural Information Processing Systems*, vol. 12, 1999.

[69] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1928–1937.

[70] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1861–1870.

[71] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *International Conference on Learning Representations (ICLR)*, 2016.

[72] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning*. PMLR, 2014, pp. 387–395.

[73] F.-M. Luo, T. Xu, H. Lai, X.-H. Chen, W. Zhang, and Y. Yu, "A survey on model-based reinforcement learning," *arXiv preprint arXiv:2206.09328*, 2022.

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

[75] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[76] G. Chaslot, S. Bakkes, I. Szita, and P. Spronck, "Monte-carlo tree search: A new framework for game ai," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 4, no. 1, 2008, pp. 216–217.

[77] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[78] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[79] R. S. Sutton, "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in *Machine learning proceedings 1990*. Elsevier, 1990, pp. 216–224.

[80] T. Kurutach, I. Clavera, Y. Duan, A. Tamar, and P. Abbeel, "Model-ensemble trust-region policy optimization," in *International Conference on Learning Representations*, 2018.

[81] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[82] S. Levine and V. Koltun, "Guided policy search," in *International Conference on Machine Learning*. PMLR, 2013, pp. 1–9.

[83] M. Deisenroth and C. E. Rasmussen, "Pilco: A model-based and data-efficient approach to policy search," in *Proceedings of the 28th International Conference on Machine Learning*. Citeseer, 2011, pp. 465–472.

[84] T. Yu, G. Thomas, L. Yu, S. Ermon, J. Y. Zou, S. Levine, C. Finn, and T. Ma, "Mopo: Model-based offline policy optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 129–14 142, 2020.

[85] Y. Chen, L. Xu, C. Gulcehre, T. Le Paine, A. Gretton, N. de Freitas, and A. Doucet, "On instrumental variable regression for deep offline policy evaluation," *Journal of Machine Learning Research*, vol. 23, no. 302, pp. 1–40, 2022.

[86] L. Liao, Z. Fu, Z. Yang, Y. Wang, M. Kolar, and Z. Wang, "Instrumental variable value iteration for causal offline reinforcement learning," *arXiv preprint arXiv:2102.09907*, 2021.

[87] D. Zhu, L. E. Li, and M. Elhoseiny, "Causaldyna: Improving generalization of dyna-style reinforcement learning via counterfactual-based data augmentation," 2021.

[88] C. Lu, B. Huang, K. Wang, J. M. Hernández-Lobato, K. Zhang, and B. Schölkopf, "Sample-efficient reinforcement learning via counterfactual-based data augmentation," in *Offline Reinforcement Learning Workshop at Neural Information Processing Systems*, 2020.

[89] Z.-M. Zhu, S. Jiang, Y.-R. Liu, Y. Yu, and K. Zhang, "Invariant action effect model for reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 9260–9268.

[90] J. Zhang and E. Bareinboim, "Markov decision processes with unobserved confounders: A causal approach," Technical report, Technical Report R-23, Purdue AI Lab, Tech. Rep., 2016.

[91] L. Wang, Z. Yang, and Z. Wang, "Provably efficient causal reinforcement learning with confounded observational data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 164–21 175, 2021.

[92] D. A. Bruns-Smith, "Model-free and model-based policy evaluation when causality is uncertain," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1116–1126.

[93] C. Shi, J. Zhu, Y. Shen, S. Luo, H. Zhu, and R. Song, "Off-policy confidence interval estimation with confounded markov decision process," *arXiv preprint arXiv:2202.10589*, 2022.

[94] H. Namkoong, R. Keramati, S. Yadlowsky, and E. Brunskill, "Off-policy policy evaluation for sequential decisions under unobserved confounding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 819–18 831, 2020.

[95] J. Guo, M. Gong, and D. Tao, "A relational intervention approach for unsupervised dynamics generalization in model-based reinforcement learning," in *International Conference on Learning Representations*, 2022.

[96] L. Buesing, T. Weber, Y. Zwols, N. Heess, S. Racaniere, A. Guez, and J.-B. Lespiau, "Woulda, coulda, shoulda: Counterfactually-guided policy search," in *International Conference on Learning Representations*, 2018.

[97] M. Oberst and D. Sontag, "Counterfactual off-policy evaluation with gumbel-max structural causal models," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4881–4890.

[98] T. W. Killian, M. Ghassemi, and S. Joshi, "Counterfactually guided policy transfer in clinical settings," in *Conference on Health, Inference, and Learning*. PMLR, 2022, pp. 5–31.

[99] G. Tennenholtz, U. Shalit, and S. Mannor, "Off-policy evaluation in partially observable environments," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 06, 2020, pp. 10 276–10 283.

[100] A. Bennett and N. Kallus, "Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes," *arXiv preprint arXiv:2110.15332*, 2021.

[101] Y. Hu and S. Wager, "Off-policy evaluation in partially observed markov decision processes," *arXiv preprint arXiv:2110.12343*, 2022.

[102] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[103] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, and N. De Freitas, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3040–3049.

[104] S. L. Barton, N. R. Waytowich, E. Zaroukian, and D. E. Asher, "Measuring collaborative emergent behavior in multi-agent reinforcement learning," in *International Conference on Human Systems Engineering and Design: Future Trends and Applications*. Springer, 2018, pp. 422–427.

[105] C. Shi, M. Uehara, J. Huang, and N. Jiang, "A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes," in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 057–20 094.

[106] A. Forney, J. Pearl, and E. Bareinboim, "Counterfactual data-fusion for online reinforcement learners," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1156–1164.

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

[108] A. Yabe, D. Hatano, H. Sumita, S. Ito, N. Kakimura, T. Fukunaga, and K.-i. Kawarabayashi, "Causal bandits with propagating inference," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5512–5520.

[109] F. Lattimore, T. Lattimore, and M. D. Reid, "Causal bandits: Learning good interventions via causal inference," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

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

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

[112] J. Zhang and E. Bareinboim, "Transfer learning in multi-armed bandit: a causal approach," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 1340–1346.

[113] S. Lee and E. Bareinboim, "Structural causal bandits: where to intervene?" *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[114] ——, "Structural causal bandits with non-manipulable variables," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019.

[115] C. Subramanian and B. Ravindran, "Causal contextual bandits with targeted interventions," in *International Conference on Learning Representations*, 2022.

[116] A. Bennett and N. Kallus, "Policy evaluation with latent confounders via optimal balance," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[117] N. Kallus and A. Zhou, "Confounding-robust policy improvement," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[118] ——, "Minimax-optimal policy learning under unobserved confounding," *Management Science*, vol. 67, no. 5, pp. 2870–2890, 2021.

[119] J. Zhang and E. Bareinboim, "Near-optimal reinforcement learning in dynamic treatment regimes," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[120] ——, "Online reinforcement learning for mixed policy scopes," *Columbia CausalAI Laboratory, Technical Report (R-84)*, 2022.

[121] S. Chen and B. Zhang, "Estimating and improving dynamic treatment regimes with a time-varying instrumental variable," *arXiv preprint arXiv:2104.07822*, 2021.

[122] J. Zhang, D. Kumor, and E. Bareinboim, "Causal imitation learning with unobserved confounders," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 263–12 274, 2020.

[123] J. Etesami and P. Geiger, "Causal transfer for imitation learning and decision making under sensor-shift," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 06, 2020, pp. 10 118–10 125.

[124] D. Kumor, J. Zhang, and E. Bareinboim, "Sequential causal imitation learning with unobserved confounders," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 669–14 680, 2021.

[125] G. Tennenholtz, A. Hallak, G. Dalal, S. Mannor, G. Chechik, and U. Shalit, "On covariate shift of latent confounders in imitation and reinforcement learning," in *International Conference on Learning Representations*, 2022.

[126] G. Swamy, S. Choudhury, J. A. Bagnell, and Z. S. Wu, "Causal imitation learning under temporally correlated noise," in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 877–20 890.

[127] Z. Zhou, M. Bloem, and N. Bambos, "Infinite time horizon maximum causal entropy inverse reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 63, no. 9, pp. 2787–2802, 2017.

[128] N. Kallus and A. Zhou, "Confounding-robust policy evaluation in infinite-horizon reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 293–22 304, 2020.

[129] S. Nair, Y. Zhu, S. Savarese, and L. Fei-Fei, "Causal induction from visual observations for goal directed tasks," *arXiv preprint arXiv:1910.01751*, 2019.

[130] I. Feliciano-Avelino, A. Méndez-Molina, E. F. Morales, and L. E. Sucar, "Causal based action selection policy for reinforcement learning," in *Mexican International Conference on Artificial Intelligence*. Springer, 2021, pp. 213–227.

[131] Y. Lu, A. Meisami, and A. Tewari, "Efficient reinforcement learning with prior causal knowledge," in *First Conference on Causal Learning and Reasoning*, 2022.

[132] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 03, 2020, pp. 2493–2500.

[133] A. Bennett, N. Kallus, L. Li, and A. Mousavi, "Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1999–2007.

[134] Y. Nair and N. Jiang, "A spectral approach to off-policy evaluation for pomdps," *The 38th International Conference on Machine Learning*, 2021.

[135] V. Nair, V. Patil, and G. Sinha, "Budgeted and non-budgeted causal bandits," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2017–2025.

[136] S. Lee and E. Bareinboim, "Characterizing optimal mixed policies: Where to intervene and what to observe," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8565–8576, 2020.

[137] M. Gonzalez-Soto, L. E. Sucar, and H. J. Escalante, "Playing against nature: causal discovery for decision making under uncertainty," *CausalML Workshop at the 35th International Conference on Machine Learning*, 2018.

[138] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

[139] T. Osa, J. Pajarinen, G. Neumann, J. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *Foundations and Trends in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[140] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, "Modeling interaction via the principle of maximum causal entropy," in *International Conference on Machine Learning*, 2010.

[141] ——, "The principle of maximum causal entropy for estimating interacting processes," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 1966–1980, 2013.

[142] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[143] Z.-M. Zhu, X.-H. Chen, H.-L. Tian, K. Zhang, and Y. Yu, "Offline reinforcement learning with causal structured world models," *arXiv preprint arXiv:2206.01474*, 2022.

[144] T. Kurutach, A. Tamar, G. Yang, S. J. Russell, and P. Abbeel, "Learning plannable representations with causal infogan," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[145] T. Herlau and R. Larsen, "Reinforcement learning of causal variables using mediation analysis," in *36th AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, 2021.

[146] A. Zhang, R. T. McAllister, R. Calandra, Y. Gal, and S. Levine, "Learning invariant representations for reinforcement learning without reconstruction," in *International Conference on Learning Representations*, 2021.

[147] M. Mozifian, A. Zhang, J. Pineau, and D. Meger, "Intervention design for effective sim2real transfer," *arXiv preprint arXiv:2012.02055*, 2020.

[148] S. Pitis, E. Creager, and A. Garg, "Counterfactual data augmentation using locally factored dynamics," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3976–3990, 2020.

[149] M. Seitzer, B. Schölkopf, and G. Martius, "Causal influence detection for improving efficiency in reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 905–22 918, 2021.

[150] F. Feng, B. Huang, K. Zhang, and S. Magliacane, "Factored adaptation for non-stationary reinforcement learning," *arXiv preprint arXiv:2203.16582*, 2022.

[151] S. Peng, X. Hu, R. Zhang, K. Tang, J. Guo, Q. Yi, R. Chen, X. Zhang, Z. Du, L. Li, Q. Guo, and Y. Chen, "Causality-driven hierarchical structure discovery for reinforcement learning," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022.

[152] Z. Wang, X. Xiao, Z. Xu, Y. Zhu, and P. Stone, "Causal dynamics learning for task-independent state abstraction," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 151–23 180.

[153] S. A. Sontakke, A. Mehrjou, L. Itti, and B. Schölkopf, "Causal curiosity: Rl agents discovering self-supervised experiments for causal representation learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9848–9858.

[154] C. Lu, B. Schölkopf, and J. M. Hernández-Lobato, "Deconfounding reinforcement learning in observational settings," *arXiv preprint arXiv:1812.10576*, 2018.

[155] B. Huang, C. Lu, L. Leqi, J. M. Hernández-Lobato, C. Glymour, B. Schölkopf, and K. Zhang, "Action-sufficient state representation learning for control with structural constraints," in *International Conference on Machine Learning*. International Machine Learning Society, 2022.

[156] B. Huang, F. Feng, C. Lu, S. Magliacane, and K. Zhang, "Adarl: What, where, and how to adapt in transfer reinforcement learning," *International Conference on Learning Representations*, 2022.

[157] A. Zhang, Z. C. Lipton, L. Pineda, K. Azizzadenesheli, A. Anandkumar, L. Itti, J. Pineau, and T. Furlanello, "Learning causal state representations of partially observable environments," *The Multi-disciplinary Conference on Reinforcement Learning and Decision Making*, 2019.

[158] A. Sonar, V. Pacelli, and A. Majumdar, "Invariant policy optimization: Towards stronger generalization in reinforcement learning," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 21–33.

[159] Y. Lu, A. Meisami, and A. Tewari, "Causal bandits with unknown graph structure," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[160] S. Saengkyongam, N. Thams, J. Peters, and N. Pfister, "Invariant policy learning: A causal perspective," *Workshop on Reinforcement Learning Theory at the 38th International Conference on Machine Learning*, 2021.

[161] G. Tennenholtz, U. Shalit, S. Mannor, and Y. Efroni, "Bandits with partially observable confounded data," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 430–439.

[162] G. E. Katz, "A cognitive robotic imitation learning system based on cause-effect reasoning," Ph.D. dissertation, University of Maryland, College Park, 2017.

[163] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[164] I. Bica, D. Jarrett, A. Hüyük, and M. van der Schaar, "Learning" what-if" explanations for sequential decision-making," *International Conference on Learning Representations*, 2021.

[165] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," *International Conference on Computer Vision(ICCV)*, 2019.

[166] I. Bica, D. Jarrett, and M. van der Schaar, "Invariant causal imitation learning for generalizable policies," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[167] G. Katz, D.-W. Huang, T. Hauge, R. Gentili, and J. Reggia, "A novel parsimonious cause-effect reasoning algorithm for robot imitation and plan recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 177–193, 2017.

[168] C. Wen, J. Lin, T. Darrell, D. Jayaraman, and Y. Gao, "Fighting copycat agents in behavioral cloning from observation histories," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2564–2575, 2020.

[169] C.-C. Chuang, D. Yang, C. Wen, and Y. Gao, "Resolving copycat problems in visual imitation learning via residual action prediction," in *17th European Conference on Computer Vision*. Springer, 2022, pp. 392–409.

[170] C. Wen, J. Lin, J. Qian, Y. Gao, and D. Jayaraman, "Keyframe-focused visual imitation learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 123–11 133.

[171] C. Wen, J. Qian, J. Lin, J. Teng, D. Jayaraman, and Y. Gao, "Fighting fire with fire: avoiding dnn shortcuts through priming," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 723–23 750.

[172] A. Méndez-Molina, "Combining reinforcement learning and causal models for robotics applications," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 8 2021, pp. 4905–4906. [Online]. Available: https://doi.org/10.24963/ijcai.2021/684

[173] W. Ding, H. Lin, B. Li, and D. Zhao, "Generalizing goal-conditioned reinforcement learning with variational causal reasoning," in *Advances in Neural Information Processing Systems*, 2022.

[174] S. Sohn, J. Oh, and H. Lee, "Hierarchical reinforcement learning for zero-shot generalization with subtask dependencies," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[175] R. Chen, X. Wu, Y. Pan, K. Yuan, L. Li, T. Ma, J. Liang, R. Zhang, K. Wang, C. Zhang *et al.*, "Eden: A unified environment framework for booming reinforcement learning algorithms," *arXiv preprint arXiv:2109.01768*, 2021.

[176] V. Luczkow, *Structural Causal Models for Reinforcement Learning*. McGill University (Canada), 2021.

[177] A. Zhang, C. Lyle, S. Sodhani, A. Filos, M. Kwiatkowska, J. Pineau, Y. Gal, and D. Precup, "Invariant causal prediction for block mdps," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 214–11 224.

[178] M. Tomar, A. Zhang, R. Calandra, M. E. Taylor, and J. Pineau, "Model-invariant state abstractions for model-based reinforcement learning," in *Self-Supervision for Reinforcement Learning Workshop - ICLR 2021*, 2021.

[179] C. Lyle, A. Zhang, M. Jiang, J. Pineau, and Y. Gal, "Resolving causal confusion in reinforcement learning via robust exploration," in *Self-Supervision for Reinforcement Learning Workshop of ICLR*, 2021.

[180] I. Dasgupta, J. Wang, S. Chiappa, J. Mitrovic, P. Ortega, D. Raposo, E. Hughes, P. Battaglia, M. Botvinick, and Z. Kurth-Nelson, "Causal reasoning from meta-reinforcement learning," *arXiv preprint arXiv:1901.08162*, 2019.

[181] D. J. Rezende, I. Danihelka, G. Papamakarios, N. R. Ke, R. Jiang, T. Weber, K. Gregor, H. Merzic, F. Viola, J. Wang *et al.*, "Causally correct partial models for reinforcement learning," *arXiv preprint arXiv:2002.02836*, 2020.

[182] Y. Sun, K. Zhang, and C. Sun, "Model-based transfer reinforcement learning based on graphical model representations," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[183] K. Kansky, T. Silver, D. A. Mély, M. Eldawy, M. Lázaro-Gredilla, X. Lou, N. Dorfman, S. Sidor, S. Phoenix, and D. George, "Schema networks: Zero-shot transfer with a generative causal model of intuitive physics," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1809–1818.

[184] M. Gasse, D. Grasset, G. Gaudron, and P.-Y. Oudeyer, "Causal reinforcement learning using observational and interventional data," *arXiv preprint arXiv:2106.14421*, 2021.

[185] J. Liang and A. Boularias, "Learning transition models with time-delayed causal relations," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8087–8093.

[186] ——, "Inferring time-delayed causal relations in pomdps from the principle of independence of cause and mechanism," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[187] N. Thams, S. Saengkyongam, N. Pfister, and J. Peters, "Statistical testing under distributional shifts," *arXiv preprint arXiv:2105.10821*, 2021.

[188] S. Volodin, N. Wichers, and J. Nixon, "Resolving spurious correlations in causal models of environments via interventions," *Causal Learning for Decision Making Workshop of ICLR*, 2020.

[189] T. E. Lee, J. A. Zhao, A. S. Sawhney, S. Girdhar, and O. Kroemer, "Causal reasoning in simulation for structure and transfer learning of robot manipulation policies," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4776–4782.

[190] C. Lu, J. M. Hernández-Lobato, and B. Schölkopf, "Invariant causal representation learning for generalization in imitation and reinforcement learning," in *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.

[191] G. Katz, D.-W. Huang, R. Gentili, and J. Reggia, "Imitation learning as cause-effect reasoning," in *International Conference on Artificial General Intelligence*. Springer, 2016, pp. 64–73.

[192] O. Ahmed, F. Träuble, A. Goyal, A. Neitz, M. Wuthrich, Y. Bengio, B. Schölkopf, and S. Bauer, "Causalworld: A robotic manipulation benchmark for causal structure and transfer learning," in *International Conference on Learning Representations*, 2020.

[193] M. P. Deisenroth, G. Neumann, J. Peters *et al.*, "A survey on policy search for robotics," *Foundations and Trends® in Robotics*, vol. 2, no. 1–2, pp. 1–142, 2013.

[194] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[195] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[196] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[197] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[198] Z. Ye, K. Xiao, Y. Ge, and Y. Deng, "Applying simulated annealing and parallel computing to the mobile sequential recommendation," *IEEE Transactions on Knowledge & Data Engineering*, vol. 31, no. 02, pp. 243–256, 2019.

[199] J.-C. Shi, Y. Yu, Q. Da, S.-Y. Chen, and A.-X. Zeng, "Virtual-taobao: Virtualizing real-world online retail environment for reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4902–4909.

[200] W. Shang, Y. Yu, Q. Li, Z. Qin, Y. Meng, and J. Ye, "Environment reconstruction with hidden confounders for reinforcement learning based recommendation," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 566–576.

[201] S. Weichwald, S. W. Mogensen, T. E. Lee, D. Baumann, O. Kroemer, I. Guyon, S. Trimpe, J. Peters, and N. Pfister, "Learning by doing: Controlling a dynamical system using causality, control, and reinforcement learning," *arXiv preprint arXiv:2202.06052*, 2022.

[202] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," *arXiv preprint arXiv:1708.05866*, 2017.

[203] A. Pritzel, B. Uria, S. Srinivasan, A. P. Badia, O. Vinyals, D. Hassabis, D. Wierstra, and C. Blundell, "Neural episodic control," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2827–2836.

[204] Wikipedia contributors, "Cognition," 2022, accessed 18-November-2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Cognition&oldid=1118273677

[205] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[206] M. Zhou, J. Luo, J. Villella, Y. Yang, D. Rusu, J. Miao, W. Zhang, M. Alban, I. FADAKAR, Z. Chen *et al.*, "Smarts: An open-source scalable multi-agent rl training school for autonomous driving," in *Conference on Robot Learning*. PMLR, 2021, pp. 264–285.

Yan Zeng received her B.S. degree in the School of Applied Mathematics in 2016 and Ph.D. degree in the School of Computer at Guangdong University of Technology in 2021. She is currently a Postdoctoral Researcher in the Department of Computer Science and Technology, Tsinghua University. She was an intern in the Causal Inference Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan from 2019 to 2020. Her current research interests include causal reinforcement learning, causal discovery, and latent variable modeling, etc.

Ruichu Cai received his B.S. in Applied Mathematics and Ph.D. in Computer Science from South China University of Technology in 2005 and 2010, respectively. He is currently a Professor in the School of Computer, Guangdong University of Technology and in the Pazhou Lab., Guangzhou. He was a visiting student at National University of Singapore in 2007-2009, and research fellow at the Advanced Digital Sciences Center, Illinois at Singapore in 2013-2014. His research interests cover a variety of different topics including causality, machine learning and their applications.

Fuchun Sun , IEEE Fellow, received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 1997. He is currently a Professor with the Department of Computer Science and Technology and President of Academic Committee of the Department, Tsinghua University, deputy director of State Key Lab. of Intelligent Technology and Systems, Beijing, China. His research interests include artificial intelligence, intelligent control and robotics, information sensing and processing in artificial cognitive systems, etc. He was recognized as a Distinguished Young Scholar in 2006 by the Natural Science Foundation of China. He became a member of the Technical Committee on Intelligent Control of the IEEE Control System Society in 2006. He serves as Editor-in-Chief of International Journal on Cognitive Computation and Systems, and an Associate Editor for a series of international journals including the IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, the IEEE TRANSACTIONS ON FUZZY SYSTEMS, and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS.

Libo Huang received his B.S. degree in Mathematics at Jiangxi Normal University in 2016 and his Ph.D. in the School of Information Engineering at Guangdong University of Technology in 2021. He was a joint Ph.D. student in the Department of Electronics and Computer Engineering at Brunel University London from 2019 to 2020. He is currently a Postdoctoral Researcher at the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include deep continuous learning and causal reinforcement learning.

Zhifeng Hao received the B.S. degree in Mathematics from the Sun Yat-Sen University in 1990, and the Ph.D. degree in Mathematics from Nanjing University in 1995. He is currently a Professor in the College of Science, Shantou University. His research interests involve various aspects of Algebra, Machine Learning, Data Mining, Evolutionary Algorithms.