Teaching Agents to Understand Cause and Effect: A Survey of Causal Reinforcement Learning with Applications

A thesis submitted to
Indian Institute of Science Education and Research Pune
in partial fulfilment of the requirements for the
Mathematics M.Sc Degree Program
under the supervision of
Dr. Anindya Goswami

by
Nilay Nitnaware
April, 2025

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road, Pashan, Pune India 411008

Certificate

This is to certify that this thesis entitled "*Teaching Agents to Understand Cause and Effect: A Survey of Causal Reinforcement Learning with Applications*" submitted towards the partial fulfilment of the Mathematics M.Sc Degree Program at the Indian Institute of Science Education and Research Pune represents work carried out by *Nilay Nitnaware* under the supervision of *Dr. Anindya Goswami*.

Dr. Anindya Goswami
Master's Thesis Supervisor

Declaration

I hereby declare that the matter embodied in the report entitled "Teaching Agents to Understand Cause and Effect: A Survey of Causal Reinforcement Learning with Applications" are the results of the work carried out by me at the Department of Mathematics, Indian Institute of Science Education and Research, Pune, under the supervision of Anindya Goswami and the same has not been submitted elsewhere for any other degree.

Nilay Nitnaware

Acknowledgment

First and foremost, I would like to express my deepest gratitude to Professor Anindya for his unwavering patience over the past three semesters and for consistently inspiring me to pursue further research. No matter how difficult life became, our Friday meetings were a beacon of hope and motivation. Week after week, those sessions lifted my spirits and rekindled my passion for learning and discovery.

I am also profoundly thankful to my friends; Drishti (Huntu) Phukon, Kwanit (Plus One) Gangopadhyay, Shinjini (Szeth) Paul and Trisha (Puku) Mistri, and . Their dedication to their own work and the love they show for their craft have been a constant source of inspiration. I am especially grateful for their unwavering support during the "November Incident", a time when I was close to losing my path. Without their encouragement and care, this thesis would never have been completed.

Finally, I owe the deepest thanks to my parents for believing in me, even when I struggled to believe in myself. Their love and support carried me through the darkest times, and for that, I am eternally grateful.

Contents

1 Introduction

A central goal of Artificial Intelligence (AI) is to develop rational agents that operate in an environment to maximize a performance measure, using prior knowledge and sensory inputs. These agents follow *policies*, which are decision rules mapping observations and history to actions. When system dynamics are known, planning algorithms can be used to compute optimal policies. However, real-world environments are often only partially known, necessitating learning from interaction.

Reinforcement Learning (RL) is a dominant framework for learning optimal decisions under uncertainty. RL includes *off-policy learning* from pre-collected data and *online learning* through real-time interaction. Most RL approaches rely on the formalism of Markov Decision Processes (MDPs), where a finite set of state variables is sufficient to encode relevant history. Extensions include multi-armed bandits, contextual bandits, and partially observable MDPs (POMDPs).

A core challenge in RL is the *curse of dimensionality*—the exponential growth of state-action spaces. Fortunately, real-world systems are often modular. Hierarchical RL exploits this modularity by decomposing tasks into simpler subtasks, significantly reducing computational complexity. The integration of deep learning has further automated representation learning in RL, as seen in systems like Deep Q-Networks, AlphaGo, and Watson.

Despite these advances, current AI systems are fragile, data-hungry, and often opaque. They struggle with robustness, generalization, and interpretability. This fragility stems from a lack of formal representation of causal structures governing the environment.

Causal Inference (CI) provides a principled framework for reasoning about interventions and counterfactuals—hypothetical outcomes under alternative actions—even when data is limited or the environment is only partially observed. Recent developments in both empirical sciences and machine learning show how causal knowledge can support principled policy evaluation and robust decision-making. However, the integration of CI into AI remains in its early stages and requires substantial theoretical and empirical progress.

To develop truly intelligent and reliable agents, AI must transition from heuristic-driven approaches to a deeper understanding of the causal principles that underpin robust decision-making under uncertainty.

2 Background

2.1 Reinforcement Learning

Reinforcement learning (RL) refers to the process by which an agent learns how to map situations to actions with the aim of maximizing a cumulative reward signal. Unlike supervised learning, the agent is not explicitly instructed on which actions to take. Instead, it must explore and discover effective strategies through trial and error. A defining challenge in RL is that actions influence not only immediate rewards but also future states and rewards, introducing the notion of delayed consequences. These two aspects trial and error search and delayed reward—are the key characteristics that distinguish RL from other forms of machine learning.

In contrast, supervised learning involves learning from a dataset of labeled examples provided by an external expert. Each training example specifies the correct response for a given input, allowing the learner to generalize its predictions to unseen instances. While effective in many domains, supervised learning is limited in interactive settings, where collecting representative and correct examples for every possible scenario is impractical. In such settings, agents must learn directly from their own experience.

RL also differs from unsupervised learning, which primarily seeks to uncover hidden patterns or structure within unlabeled data. While RL does not rely on labeled examples, its primary goal is not structural discovery but rather the maximization of expected reward over time. Although identifying structure in the environment can assist in learning more effective policies, it is not the central objective of RL. Thus, reinforcement learning stands as a distinct paradigm within machine learning—complementing, but not reducible to, either supervised or unsupervised learning.

We shall define the learning problem and ways to solve it using causal inference in upcoming sections. Primarily we'll see better solutions to the traditional tasks, i.e. Online and off-policy learning. And then we shall see 4 new causal RL tasks with expand the boundaries of RL.

2.2 Causal Inference

2.2.1 The Environment and Structural Causal Models

Definition 1 (Structural Causal Model)
A Structural Causal Model *(SCM)* $\mathcal{M}$ *is defined as a 4-tuple* (Π, V, F, P) *where:*

- $\mathbf{U} = \{U_1, U_2, \ldots, U_k\}$ *is a set of exogenous (background) variables. These represent influences external to the model that nonetheless affect variables within it.*
- $\mathbf{V} = \{V_1, V_2, \ldots, V_n\}$ *is a set of endogenous variables, which are determined by variables and functions within the model.*
- $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ *is a set of structural functions, one for each endogenous variable. Each function* f_i *defines the value of* V_i *as*

  $$v_i \coloneqq f_i(pa_i, u_i),$$

  where $pa_i \subseteq V \setminus \{V_i\}$ *are the parents of* V_i *and* $u_i \in \mathbf{U}$ *is the corresponding exogenous input.*
- $P(\mathbf{U})$ *is a probability distribution over the exogenous variables.*

Several insights emerge from the definition of a Structural Causal Model (SCM). First, an SCM $\mathcal{M}$ distinguishes between two types of variables: the exogenous variables Π, which are unobserved and determined outside the model's domain, and the endogenous variables V, which are observed and influenced within the model. The exogenous variables follow a probability distribution $P(\mathbf{U})$ over their possible states, representing factors external to the agent's deployed environment.

In practical terms, $\mathbf{U}$ often includes latent or unmeasured influences, such as a patient's genetic makeup, a customer's private demographics, or a robot's hidden physical constraints, which can significantly impact outcomes yet remain inaccessible to the agent. For example, a doctor may lack information about a patient's DNA while prescribing medication, even though it critically affects treatment efficacy. In contrast, another physician with access to such genetic data could make more informed decisions. Similarly, a robot with limited sensors may struggle to localize itself precisely, while others with better equipment achieve near-perfect awareness. These are simply different representations of reality, shaped by each agent's sensory and interventional capabilities within their respective environments.

More broadly, the existence of $\mathbf{U}$ enables the modeling of unobserved confounders, which are intrinsic to any real-world scenario where complete information is unattainable. This is especially true in human-centered environments, where full observability is inherently restricted. Unless stated otherwise, we will adopt the viewpoint of a Causal Reinforcement Learning (CRL) agent, reflecting the perceptual and interventional limitations specific to that agent.

Secondly, each endogenous variable $V_i \in V$ is determined by a structural function f_i that depends on a subset of other endogenous variables and relevant exogenous inputs. The randomness in $\mathbf{U}$ induces variability in V, laying the foundation for the formal data-generating process described in the next section. The set of structural functions $\mathcal{F}$ and the distribution $P(\mathbf{U})$ jointly specify how observable states and rewards are generated in the environment.

A key advantage of the SCM framework lies in its expressive power: it can incorporate a wide range of causal relationships. This generality allows SCMs to encompass familiar reinforcement learning environments such as Multi-Armed Bandits (MABs) and Markov Decision Processes (MDPs). A crucial distinction, however, is that standard RL models typically leave action variables X unspecified within the environment. In contrast, SCMs explicitly encode the mechanisms governing actions, which might originate from a human demonstrator, another agent, or even natural forces like gravity. We denote these mechanisms as $f_X = \{f_X : \pi_X \to X\}$, representing the functions that determine the values of X.

By default, the environment is assumed to operate under a behavior policy terminology standard in RL [4] or under the *natural regime* in the context of causal inference [3]. For instance, a human doctor chooses treatments based on their expertise, irrespective of an AI agent's recommendation. Similarly, the motion of a particle may evolve under gravitational influence, even without direct intervention. The essential point for CRL is that the agent does not inherently control how X is instantiated in the environment—this must be explicitly modeled or learned.

A few key observations are in order. First, Structural Causal Models (SCMs) offer a highly expressive framework that naturally accommodates the modeling of traditional reinforcement learning (RL) environments. Second, although the SCM defines the underlying generative mechanisms of the environment, its actual structure is typically hidden from the agent. The purpose of the SCM is merely to encode the dynamics that govern the environment's behavior and the resulting Potential Causal Histories (PCHs). Third, the SCM itself is conceptually distinct from the datasets generated through interactions with the environment and from any assumptions the agent may impose on these datasets .

2.2.2 Learning through Observational and Interventional Regimes

A key characteristic of the CRL architecture outlined so far is the clear separation between the agent and the environment in which it operates. This division naturally leads to a distinction between endogenous variables (Π) and exogenous variables ($\mathbf{U}$). Notably, this separation is not fixed or universal, it is shaped by the perspective of the specific agent. In other words, the way the environment is partitioned into endogenous and exogenous components depends on the agent's perceptual and interventional reach. Furthermore, the environment may act as an intermediary for diverse forms of interaction, effectively concealing the presence of other behavior-generating agents with their own policies and perspectives, distinct from those of the CRL agent we are focusing on.

Observational Distribution

The CRL agent interacts with its environment primarily in two ways: by perceiving it through sensors ("seeing") and by intervening via actuators ("doing"). These modes of interaction are formalized through layers 1 and 2 of the PCH. When the CRL agent simply observes the unfolding of events over time, the resulting variations stem from other behavior agents present in the environment , which could be either natural phenomena or artificial entities. Since the CRL agent lacks knowledge of these agents' internal models their perceptions, intentions, and behavior policies their influence on the environment is treated as part of a passive observational process. In such settings, the CRL agent does not actively manipulate the structural causal model (SCM); instead, action variables X evolve according to the prevailing behavior or natural regime.

For any SCM $\mathcal{M}$, the set of structural functions $\mathcal{F}$ establishes a mapping from exogenous (unobserved) variables $\mathbf{U}$ to endogenous (observed) variables $\mathbf{V}$. The probability distribution $P(\mathbf{U})$ over the exogenous variables induces a joint distribution over the endogenous variables $\mathbf{V}$, denoted by $P(\mathbf{V})$, which is referred to as the *observational distribution*.

Definition 2 (Observational Distribution)
Given a structural causal model $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, the observational distribution over any set of endogenous variables $Y \subseteq \mathbf{U}$ is defined as:

$$P(y) = \sum_{u} \mathbb{1}\{Y(u) = y\}\, P(u),$$

where $Y(u)$ denotes the value of Y obtained by evaluating the structural functions $\mathcal{F}$ with exogenous variables set to u.

Interventional Distribution

In this section, we describe how an agent's interactions with the environment can be modeled when they attempt to bring about certain outcomes. Specifically, we focus on interventions that depend on prior states and actions. These are commonly known as soft or policy interventions, and we follow a framework that formalizes them rigorously.

Let $\pi_{\mathbf{X}}$ represent a sequence of decision rules $\{\pi_{\mathbf{X}} \mid \forall X \in \mathbf{X}\}$ defined over a generic set of action variables X. Each π_X is a function that selects values for X using inputs from a set of variables $\mathbf{S}_X$; in other words, $X \leftarrow \pi_X(S_X)$. For convenience, we write this as $\pi_X(X \mid S_X)$ to indicate a stochastic policy that maps values of S_X to a probability distribution over the possible values of each $X \in \mathbf{X}$.

These policies π_X are also known as adaptive treatment strategies or treatment policies in medical decision-making. They are especially useful in personalized medicine, where treatment plans are continually adjusted according to the evolving condition of a patient.

A policy specifies the actions an agent (such as a doctor or a recommendation algorithm) should take based on the state variables it observes within a structural causal model (SCM). A policy intervention, written as $\text{do}(X \leftarrow \pi_X)$ and abbreviated as $\text{do}(\pi_X)$, replaces the default behavior policy f_X of every variable $X \in X$ with the specified decision rule π_X.

We now formally define the resulting system when such a policy intervention is applied, using the concept of a submodel.

Definition 3 (Submodel)
Let $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, \boldsymbol{P} \rangle$ be a structural causal model (SCM), and let π_X be a policy defined over a set of action variables $X \subseteq V$. The submodel $\mathcal{M}_{\pi_X}$ induced by the policy π_X is another SCM, given by $\mathcal{M}_{\pi_X} = \langle U, V, \mathcal{F}_{\pi_X}, P(U) \rangle$, where the updated structural assignments F^{π_X} are defined as:

$$\mathcal{F}_{\pi_X} = \{f_V : V \setminus X\} \cup \{\pi_X : X\}.$$

This submodel reflects the effect of replacing the structural functions associated with the variables in X by the policy π_X, while leaving all other mechanisms unchanged.

In words, when the causal model $\mathcal{M}$ is subjected to an intervention specified by π_X, the structural equations governing the variables in X are replaced with the corresponding policy mechanisms defined by π_X. All other structural functions remain unchanged. The resulting model, denoted $\mathcal{M}_{\pi_X}$, captures the altered causal dynamics under the intervention.

A notable subclass of such interventions is known as *atomic interventions*, denoted $\text{do}(X \leftarrow x)$, or simply $\text{do}(x)$. These interventions assign fixed constant values to the variables in X, effectively defining the decision rules as $X \leftarrow x$ for every $X \in \mathbf{X}$.

The submodel resulting from an atomic intervention $\text{do}(x)$ applied to a model $\mathcal{M}$ is commonly denoted by $\mathcal{M}_x$. This submodel represents a specific instantiation of the broader causal system $\mathcal{M}$, which is capable of generating multiple possible worlds. In contrast, $\mathcal{M}_x$ captures a single such world shaped by the intervention.

The significance of a submodel lies in its ability to assign precise meaning to how the world would behave under new interventional scenarios. Just as a structural causal model (SCM) naturally determines observational distributions (see 2), it also provides a principled way to evaluate the outcomes resulting from interventions of the form $\text{do}(\pi_X)$, where π_X is a policy.

The effect of such an intervention on outcome variables of interest—often reward signals denoted by Y is referred to as a *potential response*. The formal notion of potential responses to atomic interventions $\text{do}(x)$ was introduced by Pearl (2000). We now extend that concept to encompass policy-based interventions $\text{do}(\pi_X)$.

Definition 4 (Potential Response)
Let $\mathcal{M} = \langle U, V, F, P \rangle$ be a structural causal model (SCM), and let $X, Y \subseteq V$ be two sets of variables. Let π_X be a policy over the variables in X, and let $u \in U$ represent a specific unit or context. The potential response of Y to the policy intervention $do(\pi_X)$, denoted $Y_{\pi_X}(u)$, is defined as the solution for Y obtained from the system of structural equations F_{π_X} in the modified model $\mathcal{M}_{\pi_X}$. That is,

$$Y_{\pi_X}(u) \coloneqq Y(u; \mathcal{M}_{\pi_X}).$$

Definition 5 (Interventional Distribution)
Let $\mathcal{M} = \langle U, V, F, P(U) \rangle$ be a structural causal model (SCM). This model induces a family of joint distributions over the endogenous variables V, each corresponding to an intervention $do(\pi_X)$, where π_X is a policy over a set of action variables $X \subseteq V$.

For any subset $Y \subseteq V$, the interventional distribution under policy π_X is defined as:

$$P^{\pi_X}(Y = y) \coloneqq \sum_{u : Y^{\pi_X}(u) = y} P(u),$$

where $Y^{\pi_X}(u)$ denotes the potential response of Y under the policy intervention $do(\pi_X)$ (as given in Definition 4).

This distribution can be computed via the following steps:

1. *Substitute the original mechanisms of each variable in X with the corresponding policy functions π_X, resulting in the submodel $\mathcal{M}^{\pi_X}$;*

2. *For each context $u \in U$, evaluate the structural equations F^{π_X} following a valid topological order (ensuring variables on the left-hand side are evaluated only after those on the right-hand side);*

3. *Accumulate the probability mass $P(U = u)$ for all u such that the evaluated outcome $Y^{\pi_X}(u)$ equals y in the submodel $\mathcal{M}^{\pi_X}$.*

Causal Decision Models

We now formalize the policy optimization problem within the framework of structural causal models (SCMs), using the causal tools developed earlier. The environment is represented by an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, \mathcal{P} \rangle$. The agent's objective is to interact with a subset of variables $X \subseteq \mathbf{V}$, considered as actions, in order to optimize performance measured by a reward variable $Y \subseteq \mathbf{V}$, evaluated within the environment $\mathcal{M}$.

The agent selects the values of the action variables X by applying interventions $\text{do}(\pi)$ according to a policy π. The set of all such candidate policies over X defines the policy space Π. We formalize this below.

Definition 6 (**Policy Space**)
Let $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, \mathcal{P} \rangle$ be a structural causal model, and let $X = \{X_1, \ldots, X_H\} \subseteq \mathbf{V}$ be a collection of action variables. A policy space Π is a set of policies π, where each policy is a sequence of decision rules

$$\pi = (\pi_1(X_1 \mid S_1), \ldots, \pi_H(X_H \mid S_H))$$

such that for every $i = 1, \ldots, H$:

- *The action variable X_i is not a descendant of any future actions $X_{i+1}, \ldots, X_H$, that is, $X_i \in \mathbf{V} \setminus De(X_{i+1}, \ldots, X_H)$;*

- *The state variables S_i are non descendants of the actions $X_i, \ldots, X_H$, that is, $S_i \subseteq \mathbf{V} \setminus De(X_i, \ldots, X_H)$.*

We will denote such a policy space compactly as:

$$\Pi = \{(X_1, S_1), \ldots, (X_H, S_H)\}$$

Every policy $\pi \in \Pi$ is described as a sequence of decision rules:

$$(\pi_1(X_1 \mid S_1), \ldots, \pi_H(X_H \mid S_H)).$$

An agent following policy π selects actions $X_1, \ldots, X_H$ in temporal order, guided by this sequence.

At each intervention step $i = 1, \ldots, H$, the agent:

1. Observes the state variables $S_i = s_i$;
2. Samples an action $x_i \sim \pi_i(X_i \mid S_i = s_i)$ based on the decision rule π_i;
3. Executes the intervention $\text{do}(X_i \leftarrow x_i)$, applying the selected action.

In summary, a policy space Π defines:

- the set of controllable actions $X = (X_1, \ldots, X_H)$, and
- the state variables $S_1, \ldots, S_H$ that are observable at the time of each intervention.

Definition 7 (Causal Decision Model)
A causal decision model *(CDM) is a tuple*

$$\langle \mathcal{M}, \Pi, R \rangle$$

where $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, \mathcal{P} \rangle$ *is a structural causal model (SCM),* Π *is a policy space defined over actions* $X \subseteq \mathbf{V}$*, and* R *is a reward function defined over reward signals* $Y \subseteq \mathbf{V}$*.*

Among the elements in Definition 7, the SCM $\mathcal{M}$ represents the underlying environment; the policy space Π indicates the agent's capabilities once deployed in the environment, i.e., which variables it can control and observe during interaction; and the reward function R evaluates the agent's performance from the system designer's perspective.

Formally, every CDM $\langle \mathcal{M}, \Pi, R \rangle$ describes a planning or decision making task (Bellman, 1966) aimed at identifying an optimal strategy from the policy space Π, given full knowledge of the environment $\mathcal{M}$. The optimal policy π^* for a CDM $\langle \mathcal{M}, \Pi, R \rangle$ is one that maximizes the reward function R under the dynamics of $\mathcal{M}$, that is:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_\pi[R(Y); \mathcal{M}]$$

Each CDM $\langle \mathcal{M}, \Pi, R \rangle$ can be visualized using an augmented causal diagram G built from the SCM $\mathcal{M}$. In this diagram, action variables X and reward variables Y are highlighted in blue and red, respectively; for every action $X_i \in X$, its input state S_i.

A few key observations:

- The number of action variables $H = |X|$ defines the horizon of the decision making problem. When $H = 1$, the CDM models a single stage decision process such as multi armed bandits. When $H > 1$, it captures a sequential decision making setting, like in Markov decision processes, where the agent must choose each action X_i based on the state S_i observed at that step.

- Different CDMs can be defined over the same environment $\mathcal{M}$ by altering the policy space Π and reward function R, yielding distinct optimal policies. That is, the optimal policy π^* depends on the agent's interactive capabilities after deployment and how its behavior is incentivized. Changing either Π or R will affect the resulting optimal policy.

More broadly, the formulation of CDMs allows one to represent canonical planning tasks (or equivalently, decision making models) across disciplines, including RL and healthcare, when the detailed parametrization of the underlying environment $\mathcal{M}$ is known.

- **Multi Armed Bandit** .
  $\langle \mathcal{M}^*{}_{MAB}, \Pi, Y \rangle$, consisting of an arm choice X and reward Y. The policy space $\Pi = \{\langle X, \cdot \rangle\}$ defines a set of policies π that selects values of action X following a probability distribution $\pi(X)$.

- **Contextual Bandit** .
  $\langle \mathcal{M}^*{}_{CMAB}, \Pi, Y \rangle$. Compared to MABs, a context variable S is now observed. The policy space $\Pi = \{\langle X, S \rangle\}$ consists of candidate policies $\pi(X|S)$ that select values of action X based on the observed context S.

- **Dynamic Treatment Regime** . In a DTR model $\langle \mathcal{M}^*_{DTR}, \Pi, Y \rangle$, the policy space $\Pi = \{\langle X_i, \{S_1, \ldots, S_i, X_1, \ldots, X_{i1}\} \rangle\}_{i=1}^{H}$ consists of a sequence of decision rules $\pi = (\pi_1, \ldots, \pi_H)$. For each i th stage, the decision rule $\pi_i(X_i|S_1, \ldots, S_i, X_1, \ldots, X_{i1})$ selects values of action X_i based on all past actions and states' history $S_1, \ldots, S_i, X_1, \ldots, X_{i1}$. The goal is to maximize the primary outcome Y after intervening on all actions $X_1, \ldots, X_H$.

- **Markov Decision Process** . Consider a Markov decision process (MDP) model $\langle \mathcal{M}, \text{MDP}, \Pi, R \rangle$. The environment $\mathcal{M}$ consists of a set of states $S = \{S_i\}_{i=1}^n$, a set of actions $X = \{X_i\}_{i=1}^n$, and a set of reward signals $Y = \{Y_i\}_{i=1}^n$. The policy space $\Pi = \{\langle X_i, S_i \rangle\}_{i=1}^n$ defines a set of decision rules $\pi = (\pi_1(X_1|S_1), \ldots, \pi_n(X_n|S_n))$. The reward function

R can be described as either the average reward $R_{\text{average}}(Y)$ or the discounted reward $R_{\text{discount}}(Y)$. In the discounted case, rewards obtained later are discounted more than those obtained earlier. If the discount factor $\vartheta = 0$, the agent is said to be myopic, i.e., only concerned with immediate rewards. Fig. 10d represents the causal diagram of an MDP model spanning over steps $i = 1, 2, 3$.

- **Partially Observable MDP** . A partially observable MDP (POMDP) is a generalization of MDP in which system dynamics are governed by an MDP environment, but the agent cannot directly utilize the underlying state $S = \{S_i\}_{i=1}^n$ to determine its actions $X = \{X_i\}_{i=1}^n$. Instead, it only receives a set of observation variables $O = \{O_i\}_{i=1}^n$ depending on the underlying states. The policy space $\Pi = \{\langle X_i, \{O_1, \ldots, O_i, X_1, \ldots, X_{i1}\}\rangle\}_{i=1}^n$ defines a set of non Markov policies $\pi = (\pi_i(X_i|O_1, \ldots, O_i, X_1, \ldots, X_{i1}))_{i=1}^n$. For each intervention stage, an agent following a non Markov policy $\pi \in \Pi$ selects an action $x_i \leftarrow \pi_i(X_i|O_1, \ldots, O_i, X_1, \ldots, X_{i1})$ based on all past observations and actions.

When the policy space Π and reward function R are well specified, and detailed parameters of the underlying environment $\mathcal{M}$ are provided, efficient algorithms exist in the planning literature to solve for an optimal policy in a CDM $\langle \mathcal{M}, \Pi, R\rangle$. For example, in an MDP model, one could obtain an optimal policy using standard dynamic programming algorithms. The same planning procedure applies to a DTR model. Planning in POMDP models is more computationally challenging due to the latent nature of the underlying states and requires the algorithm to maintain memory and possibly reason about beliefs over the states. A variety of heuristics for approximate planning in POMDPs have been proposed. Optimizing policies in a general CDM has been studied under the framework of influence diagrams, with several algorithms and approximate procedures proposed.

2.2.3 Causal Reinforcement Learning Tasks

The causal decision model described so far assumes full knowledge of the underlying environment. However, in many real world applications, the detailed parametrization of the environment is rarely known, which means that standard planning algorithms are not immediately applicable. In order for the agent to optimize the performance of the underlying system, a learning process must take place, leading to the learning paradigm of *causal reinforcement learning* (CRL).

To make this more precise, consider again a CDM $\langle \mathcal{M}, \Pi, R \rangle$, which defines a planning task of finding an optimal policy in space Π that maximizes the reward function R, evaluated within the environment $\mathcal{M}$. A CRL agent $\mathcal{C}$ is assumed to have access to the policy space Π and the performance measure R.[1] However, the underlying environment $\mathcal{M}$ is not fully known. Instead, the agent $\mathcal{C}$ only has access to:

1. a set of *structural assumptions* $\mathcal{A}$, encoding qualitative knowledge about the environment $\mathcal{M}$, and

2. a *learning regime* $\mathcal{L}$, dictating how it can interact with $\mathcal{M}$ to collect data.

This partial knowledge introduces new dimensions for formulating optimal decision making under uncertainty, which we outline below.

Learning Regimes ($\mathcal{L}$). Following the discussion of the PCH, a CRL agent may interact with $\mathcal{M}$ in different ways, such as through passive observation (i.e., $\mathcal{L} = \texttt{see}$) or active interventions (i.e., $\mathcal{L} = \texttt{do}$). These learning regimes model distinct interaction types:

- **Passive Observation.** The agent does not actively determine actions, but instead receives observational data $D \sim P(V)$ summarizing trajectories generated by another agent (e.g., a human demonstrator) already operating in the environment under some behavioral policy. This setting corresponds to off policy reinforcement learning (Li et al., 2011, 2014).

- **Active Intervention.** The agent actively selects actions X by performing interventions $\texttt{do}(\pi)$, according to a policy π, and receives interventional data $D \sim P_{\pi}(V)$. This corresponds to online reinforcement learning (Auer et al., 2002b).

We will explore these learning scenarios and their corresponding algorithms in more detail later in the paper.

Structural Assumptions ($\mathcal{A}$). Structural assumptions define a hypothesis class of plausible environmental models within which the agent operates. A common approach is to specify these assumptions via a causal diagram G, encoding the qualitative structure of the SCM $\mathcal{M}$. The hypothesis class $\mathcal{M}$ is then defined as the set of SCMs compatible with G, i.e., $\mathcal{G}(\mathcal{M}) = G$.

[1] We assume the agent is aware of how rewards are evaluated.

For instance, the causal diagram in Fig. 10d specifies a class of MDP environments consisting of states S_i, actions X_i, and reward signals Y_i for $i = 1, 2, \ldots$, while leaving the underlying transition and reward distributions unspecified. Additional structural assumptions may involve constraints on the behavior policy (Pearl and Robins, 1995), or assumptions about equivalence classes of causal diagrams (Zhang, 2008), among others.

Returning to the CDM $\langle \mathcal{M}, \Pi, R \rangle$, where the environment $\mathcal{M}$ is not fully specified, replacing $\mathcal{M}$ with the learning regime $\mathcal{L}$ and structural assumptions $\mathcal{A}$ leads to a new formulation: $\langle \mathcal{L}, \mathcal{A}, \Pi, R \rangle$. This tuple characterizes a *causal reinforcement learning task.*

Definition 8 (Causal Reinforcement Learning Task)
Let $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, \mathcal{P} \rangle$ be a structural causal model (SCM). A causal reinforcement learning (CRL) task $\mathcal{T}$ in the environment $\mathcal{M}$ is defined as a 4 tuple $\langle \mathcal{L}, \mathcal{A}, \Pi, R \rangle$, where:

1. *$\mathcal{L}$ is a learning regime describing how the agent interacts with the environment $\mathcal{M}$; for example, passive observation (`see`) or active intervention (`do`).*
2. *$\mathcal{A}$ is a set of structural assumptions about the SCM $\mathcal{M}$.*
3. *Π is a policy space over the set of actions $\mathbf{X} \subseteq \mathbf{V}$.*
4. *R is a reward function defined over reward signals $\mathbf{Y} \subseteq \mathbf{V}$.*

Each CRL task $\langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$ defines a decision making problem under uncertainty about the environment. Given this tuple, the agent aims to identify an optimal policy $\pi^* \in \Pi$ that maximizes the expected reward $\mathcal{R}$, evaluated under the true (but unknown) environment $\mathcal{M}^*$. Since $\mathcal{M}^*$ is not directly observable, the agent relies on structural assumptions $\mathcal{A}$ and a learning regime $\mathcal{L}$ to guide its interaction with the environment. The optimal policy is thus given by:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\mathbf{Y}) \mid \mathcal{A}, \mathcal{L} \right]$$

This formulation differs from Equation [ref] by expressing the unknown SCM $\mathcal{M}^*$ as a superscript, while the information accessible to the agent (i.e., $\mathcal{A}$ and $\mathcal{L}$) is placed in the conditioning set.

Algorithm 1: Causal Reinforcement Learning Agent C

Require: *CRL task* $\mathcal{T} = \langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$ ***Ensure:*** *A policy estimate* $\hat{\pi} \in \Pi$ *that optimizes the CDM* $\langle \mathcal{M}, \Pi, \mathcal{R} \rangle$

1. *Initialize dataset* $D = \{\}$
2. ***for*** *each episode* $t = 1, \ldots, T$ ***do***
   - *Interact with SCM* $\mathcal{M}$ *following regime* $\mathcal{L}$ *and receive samples* $\mathbf{V}^{(t)} \sim P_{\mathcal{L}}(\mathbf{V}; \mathcal{M})$
   - *Update dataset* $D = D \cup \{\mathbf{V}^{(t)}\}$
3. ***end for***
4. *Return an empirical estimate* $\hat{\pi} \in \Pi$ *of the optimal policy* π^* *using data* D *and assumptions* $\mathcal{A}$

A CRL task $\mathcal{T}$ in an unknown bandit environment $\mathcal{M}$, involving an arm choice X, a reward Y, and a covariate Z. The objective is to learn a policy $X \leftarrow \vec{x}$ from the policy space $\Pi = \{X, \cdot\}$ that maximizes $\mathbb{E}_x[R(Y)]$. Depending on the learning regime $\mathcal{L}$, the agent either passively observes data from $P(X, Y, Z)$ ($\mathcal{L} =$ see) or actively intervenes ($\mathcal{L} =$ do), receiving data from $P_x(Y, Z)$. Structural assumptions $\mathcal{A} = G_{\text{backdoor}}$ encode known causal relationships in the environment.

Algorithm 1 outlines the agent's learning strategy: at each episode $t = 1, \ldots, T$, it interacts with the environment to collect data $\mathbf{V}^{(t)}$ following regime $\mathcal{L}$. Observational data arises from a prior policy f_X, while interventional data follows policy π. After collecting data D, the agent computes an empirical policy estimate $\hat{\pi} \in \Pi$ using $\mathcal{A}$ and D.

This setup summarizes various policy learning problems across reinforcement learning and causal inference literature.

The formalization of CRL tasks using structural causal models provides a unifying framework for capturing a broad range of learning settings commonly studied in both reinforcement learning (RL) and causal inference (CI). Moreover, this framework allows for the systematic exploration of novel CRL tasks that arise in practical, real world scenarios.

In Section [3], we will revisit these tasks such as off policy learning, online learning, and causal identification through the lens of CRL. These tasks typically aim to optimize policies within the experimental policy space Π_{EXP} and can be viewed as variations along the first two CRL dimensions: the learning regime and structural assumptions. Other aspects of the task formulation remain fixed.

Although widely studied, some of these tasks lacked a formal understanding of key validity conditions prior to this framework. For example, verifying whether an off policy method (e.g., inverse propensity weighting or dynamic programming) can recover a policy consistent with the true optimum defined by the underlying SCM. Viewing these classical settings through the CRL framework sheds light on foundational questions and deepens our understanding of how causal inference and reinforcement learning methodologies align.

Comparison with Markov Decision Processes

The CRL tasks described so far assume that the agent has access to either the learning regime L under which the data are collected, or structural assumptions A encoding causal invariances about the environment. In this section, we will demonstrate that such causal knowledge is generally indispensable for learning an optimal policy in an unknown SCM. Our discussion focuses on standard MDPs, which represent a widely used class of sequential decision making models.

Definition 9 (Standard MDP)
A Markov decision process is a tuple $\langle \mathcal{D}(S), \mathcal{D}(X), T, R \rangle$*, where:*

1. $\mathcal{D}(S)$ *is the set of states, called the state space;*
2. $\mathcal{D}(X)$ *is the set of actions, called the action space;*
3. $T(s, x, s') \in [0, 1]$ *is the transition probability that taking action* x *in state* s *leads to state* s'*;*
4. $R(s, x)$ *is the immediate reward received after taking action* x *in state* s*.*

A policy $\pi(x \mid s)$ *is a function mapping each state in* $\mathcal{D}(S)$ *to a probability distribution over actions in* $\mathcal{D}(X)$*. For indices* $i < j \in \mathbb{N}$*, denote the sequence* $\bar{V}_{i:j} = \{V_i, V_{i+1}, \ldots, V_j\}$*.*

Given a policy π *and an initial state distribution* $P(S_1)$*, the MDP defines a joint distribution over states* $\bar{S}_{1:H}$*, actions* $\bar{X}_{1:H}$*, and rewards* $\bar{Y}_{1:H}$ *up to horizon* H*:*

$$P_\pi(s_{1:H}, x_{1:H}, y_{1:H}) = P(s_1) \prod_{i=1}^{H} \pi(x_i \mid s_i) T(s_i, x_i, s_{i+1}) \cdot \mathbb{1}\{R(s_i, x_i) = y_i\}$$

The key assumption of a standard MDP is the Markov property*, which asserts that the future is conditionally independent of the past given the present state:*

$$\bar{S}_{1:i1}, \bar{X}_{1:i1}, \bar{Y}_{1:i1} \perp\!\!\!\perp \bar{S}_{i+1:H}, \bar{X}_{i:H}, \bar{Y}_{i:H} \mid S_i \quad \textit{for all } i = 2, 3, \ldots$$

This implies the MDP is a compact representation of the joint distribution over full trajectories, provided the Markov property holds. In causal terms, the property may hold in both the observational distribution $P(s_{1:H}, x_{1:H}, y_{1:H})$ *and the interventional distribution* $P_\pi(s_{1:H}, x_{1:H}, y_{1:H})$*.*

Proposition 1 (Observational $\not\Rightarrow$ Interventional)
For any structural causal model (SCM) $\mathcal{M}^*$ *that is compatible with the causal diagram* $\mathcal{G}_{MDP}$ *in Fig. [ref], there exists another SCM* $\mathcal{M}^{(1)}$*, also compatible with* $\mathcal{G}_{MDP}$*, such that for every stage* $i = 1, 2, \ldots$*, the following equalities hold:*

$$P^{(1)}(s_{i+1} \mid s_i, x_i) = P^*(s_{i+1} \mid s_i, x_i), \quad \mathbb{E}^{(1)}[Y_i \mid s_i, x_i] = \mathbb{E}^*[Y_i \mid s_i, x_i],$$

while under the intervention on x_i*, we have:*

$$P^{(1)}_{x_i}(s_{i+1} \mid s_i) \neq P^*_{x_i}(s_{i+1} \mid s_i), \quad \mathbb{E}^{(1)}_{x_i}[Y_i \mid s_i] \neq \mathbb{E}^*_{x_i}[Y_i \mid s_i].$$

Proposition 2 (Interventional $\not\Rightarrow$ Observational)
For any structural causal model (SCM) $\mathcal{M}^*$ *that is compatible with the causal diagram* $\mathcal{G}_{MDP}$ *shown in Fig. 10d, there exists another SCM* $\mathcal{M}^{(2)}$*, also compatible with* $\mathcal{G}_{MDP}$*, such that for every stage* $i = 1, 2, \ldots$*, we have:*

$$P^{(2)}_{x_i}(s_{i+1} \mid s_i) = P^*_{x_i}(s_{i+1} \mid s_i), \quad \mathbb{E}^{(2)}_{x_i}[Y_i \mid s_i] = \mathbb{E}^*_{x_i}[Y_i \mid s_i],$$

while:

$$P^{(2)}(s_{i+1} \mid s_i, x_i) \neq P^*(s_{i+1} \mid s_i, x_i), \quad \mathbb{E}^{(2)}[Y_i \mid s_i, x_i] \neq \mathbb{E}^*[Y_i \mid s_i, x_i].$$

3 Traditional tasks through causal lens

The formal definition of causal reinforcement learning (CRL) tasks provides a unified framework for representing key problems in both reinforcement learning (RL) and causal inference (CI). In this section, we examine classical RL CI tasks, including off policy learning, online learning, and causal identification. These tasks primarily vary along two dimensions: interaction regime and structural assumptions, while other aspects remain fixed. Then we'll explores broader CRL scenarios by relaxing these constraints.

Although rooted in established literature, these tasks reveal nuanced interactions between RL and causal invariances, now made precise through the CRL lens. For example, structural causal models (SCMs) formally justify methods like inverse propensity weighting and dynamic programming in off policy learning, clarifying their applicability under unobserved confounding. This CRL perspective highlights foundational links between CI and RL.

Before delving into these tasks, we introduce additional notation. We consider a finite horizon decision process with horizon $H = |\mathbf{X}| < \infty$, where actions $\mathbf{X} = \{X_1, \ldots, X_H\}$ follow a topological order induced by the SCM $\mathcal{M}^\star$. For indices $i < j$, define the action and state sequences as $\bar{X}_{i:j} = \{X_i, \ldots, X_j\}$ and $\bar{S}_{i:j} = \{S_i, \ldots, S_j\}$, respectively. We also write $\bar{X}_i = \bar{X}_{1:i}$ and $\bar{S}_i = \bar{S}_{1:i}$. Fix a policy $\pi \in \Pi$, and for any $i \leq j$, let $(\pi_i, \ldots, \pi_j)$ denote the corresponding sequence of decision rules governing $X_i, \ldots, X_j$.

3.1 Off Policy Learning

Off policy learning addresses the challenge of estimating an optimal policy using data generated by a different behavior policy, under the key assumption of No Unmeasured Confounders (NUC), defined below. This assumption legitimizes policy learning from observational data.

We briefly introduce two major evaluation methods in off policy settings: inverse propensity weighting and dynamic programming.

The agent interacts with the environment (SCM) by passively observing sequences of events across episodes $t = 1, \ldots, T$. In each episode, the agent observes $\mathcal{M}^\star$ and samples $V(t) \sim \mathbb{P}(V)$. The agent's goal is to learn a policy from a candidate set Π that maximizes a reward function $R(Y)$. This setup is central to comparing learning frameworks and modes of interaction.

The off policy learning setting is characterized by:

$$\mathcal{T}_{\text{off}} = \left\langle I = \text{see}, \quad \mathcal{A} = \text{NUC}, \quad \Pi = \{\langle X_i, S_i \rangle\}_{i=1}^{H}, \quad \mathcal{R} : \mathcal{D}(Y) \to \mathbb{R} \right\rangle.$$

Here, the agent combines observational data with the NUC assumption, stating that there are no hidden variables influencing both actions and outcomes. The learning objective becomes:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[R(Y) \,|\, A = \text{NUC}, \; D_{\text{obs}} \sim \mathbb{P}(V) \right].$$

Practically, NUC implies that the effect of any action X_i on reward Y, conditioned on prior actions and covariates, is fully captured by observable causal pathways. No unmeasured variable should induce spurious correlation between decisions and outcomes. The following definition formalizes this assumption.

Definition 10 (No Unmeasured Confounder)
Let $\mathcal{M}^*$ be a structural causal model (SCM), and let the policy space be $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$. The No Unmeasured Confounder *(NUC) condition holds if, for each action variable X_i:*

1. *The endogenous parents $\boldsymbol{PA}_i$ are included in the observable history $\overline{\boldsymbol{X}}_{i1} \cup \overline{\boldsymbol{S}}_i$, i.e.,*
$$\boldsymbol{PA}_i \subseteq \overline{\boldsymbol{X}}_{i1} \cup \overline{\boldsymbol{S}}_i.$$

2. *The exogenous parent U_i is independent of all other exogenous noises U_j associated with endogenous variables $V_j \in V \setminus \{X_i\}$:*
$$U_i \perp\!\!\!\perp \{U_j \mid V_j \in V \setminus \{X_i\}\}.$$

In the definition above, the first condition stipulates that all endogenous parents of each action X_i are accessible, being included within the past states and actions, denoted as $\bar{S}_i, \bar{X}_{i1}$. The second condition asserts that, given the historical trajectory $\bar{\mathbf{X}}_{i1} \cup \bar{\mathbf{S}}_i$, the value of each action X_i is determined by an independent exogenous noise variable U_i.

Within the structural causal model $\mathcal{M}^*$, observational data are assumed to arise from a behavior policy f_X that assigns values to each action X_i based on its endogenous parents PA_i and the corresponding exogenous inputs U_i, across time steps $i = 1, \ldots, H$. The No Unmeasured Confounder (NUC) condition allows this behavior policy to be represented through a sequence of decision rules

$$\pi_i(X_i \mid \bar{X}_{i1}, \bar{S}_i) \quad \text{for } i = 1, \ldots, H$$

such that for each time step i, the rule satisfies $\pi_i(X_i \mid \bar{X}_{i1}, \bar{S}_i) = P(X_i \mid \text{PA}_i)$. When actions are assigned according to these rules, the following conditional independence relationships hold for any sequence of actions $\bar{x}_H$:

$$X_i \perp\!\!\!\perp S_{i+1_{\bar{x}_i}}, \ldots, S_{H_{\bar{x}_{i1}}}, Y_{\bar{x}_H} \mid \bar{X}_{i1}, \bar{S}_i \quad \text{for } i = 1, \ldots, H.$$

In this expression, $S_{i_{\bar{x}_{i1}}}$ denotes the potential outcome for the state variable in the submodel $\mathcal{M}^*_{\bar{x}_{i1}}$ induced by the intervention $\text{do}(\bar{X}_{i1} \leftarrow \bar{x}_{i1})$, and $Y_{\bar{x}_H}$ represents the prospective reward in the submodel $\mathcal{M}^*_{\bar{x}_H}$.

3.1.1 Off Policy Evaluation

When the Non Unique Confounding (NUC) assumption is satisfied, it becomes feasible to utilize various strategies to estimate and compare the outcomes of different candidate policies using observational data, without requiring direct online experimentation in the environment. One such widely used method is based on the concept of *inverse probability weighting* (IPW), a technique extensively studied and applied in the literature.

Theorem 1 (Inverse Propensity Weighting, under NUC)
Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a causal decision model (CDM), where the policy space is defined as $\Pi = \{\langle X_i, S_i \rangle\}_{i=1}^{H}$, and the reward function is $\mathcal{R} : \mathcal{D}(Y) \to \mathbb{R}$. Assuming the NUC condition holds, for any evaluation policy $\pi \in \Pi$, the expected reward under π can be computed from the observational distribution $P(V)$ using the following expression:

$$\mathbb{E}_\pi [R(Y)] = \sum_{\bar{x}_H, \bar{s}_H} \underbrace{\mathbb{E}[R(Y) \mid \bar{x}_H, \bar{s}_H] \, P(\bar{x}_H, \bar{s}_H)}_{\textit{observational distribution}} \underbrace{\prod_{i=1}^{H} \frac{\pi_i(x_i \mid s_i)}{P(x_i \mid \bar{x}_{i1}, \bar{s}_i)}}_{\textit{ratio } \pi \textit{ and obs. probabilities}} . \quad (1)$$

Among the terms presented above, the quantity $P(x_i \mid \bar{x}_{i1}, \bar{s}_i)$ represents the inherent likelihood of the behavior policy selecting action X_i at time step i, and is referred to as the *propensity score*. In order for the inverse probability weighting (IPW) estimator to be valid, it must satisfy the *positivity assumption*, which requires that the propensity scores obey

$$P(x_i \mid \bar{x}_{i1}, \bar{s}_i) > 0 \quad \text{for all combinations of } \bar{x}_i \text{ and } \bar{s}_i.$$

This condition ensures that every action under consideration by the evaluation policy has a non zero probability of appearing in the observational data.

We also observe that by successively applying Bayes' rule, the inverse probability weighting (IPW) expression in Equation (1) can be reformulated as:

$$\mathbb{E}_\pi\left[R(Y)\right] = \sum_{\bar{x}_H, \bar{s}_H} \mathbb{E}\left[R(Y) \mid \bar{x}_H, \bar{s}_H\right] \prod_{i=1}^{H} P(s_i \mid \bar{x}_{i1}, \bar{s}_{i1})\, \pi_i(x_i \mid s_i). \tag{2}$$

Evaluating this expression by traversing the actions in reverse topological order, i.e., from $i = H$ down to $i = 1$, yields an alternative method for assessing the outcomes of candidate policies. This approach is grounded in *dynamic programming* (DP). The following theorem outlines how DP can be utilized for off policy evaluation using the observational distribution, assuming the NUC condition is satisfied.

Theorem 2 (Dynamic Programming)
Let $\langle \mathcal{M}^*, \Pi, R \rangle$ be a causal decision model (CDM), where $\Pi = \{(X_i, S_i)\}_{i=1}^{H}$ and $R : \mathcal{D}(Y) \mapsto \mathbb{R}$. If the NUC condition is satisfied, then for any policy $\pi \in \Pi$, the expected reward $\mathbb{E}_\pi[R(Y)]$ can be computed from the joint distribution $P(V)$ as follows:

$$\mathbb{E}_\pi[R(Y)] = \mathbb{E}\left[\sum_{x_1} Q_\pi^{(1)}(x_1, S_1)\, \pi_1(x_1 \mid S_1)\right], \tag{3}$$

where the value function $Q_\pi^{(i)}(\bar{x}_i, \bar{s}_i)$, for $i = 1, \ldots, H1$, is defined recursively as:

$$Q_\pi^{(i)}(\bar{x}_i, \bar{s}_i) = \mathbb{E}\left[\sum_{x_{i+1}} Q_\pi^{(i+1)}(\bar{x}_{i+1}, \bar{s}_i, S_{i+1})\, \pi_{i+1}(x_{i+1} \mid S_{i+1}) \,\middle|\, \bar{x}_i, \bar{s}_i\right], \tag{4}$$

and the terminal value is given by:

$$Q_\pi^{(H)}(\bar{x}_H, \bar{s}_H) = \mathbb{E}\left[R(Y) \,|\, \bar{x}_H, \bar{s}_H\right]. \tag{5}$$

Once the IPW and DP evaluation formulations are established, practical algorithms can estimate the expected rewards of candidate policies using finite samples from the observational distribution $P(V)$.

For IPW, each observed reward is weighted by the odds ratio between the target policy π and the behavior policy's propensity score $P(x_i \mid \bar{x}_{i11}, \bar{s}_i)$ corresponding to the ratio. The expected reward is then approximated by the empirical average of these weighted rewards. Initially developed for single stage decisions ($H = 1$), this method has been extended to the sequential setting ($H > 1$), provided the NUC condition holds.

For DP based evaluation, one first approximates the value function Q using parametric models fitted to the observational data, such as linear functions via least squares regression or more flexible alternatives like trees, kernels, or neural networks. This approach falls under the broader scope of batch reinforcement learning.

Once the expected rewards are estimable, the agent may optimize over the policy space Π to obtain an optimal policy via methods like policy gradient. If the Q function $Q_\pi^{(i)}(\bar{x}_i, \bar{s}_i)$ depends only on the current state action pair (x_i, s_i), an optimal policy can be derived by iteratively optimizing each decision rule π_i in reverse topological order. This process closely resembles the structure of Q learning.

However, both IPW and DP approaches may fail when the NUC condition (Definition 10) is violated.

3.2 Online Learning

In online learning, an agent evaluates candidate policies $\pi \in \Pi$ by actively deploying them in the environment across multiple episodes $t = 1, \ldots, T$. During each episode, the agent selects a policy $\pi^{(t)}$, intervenes on the action variables via $(X \leftarrow \pi^{(t)})$, and observes outcomes $V^{(t)} \sim P_{\pi^{(t)}}(V)$.

Formally, the online learning task is defined as:

$$\mathcal{T}_{\text{on}} = \langle \mathcal{L} = \texttt{do}, \mathcal{A} = \emptyset, \Pi = \{(X_i, S_i)\}_{i=1}^{H}, \mathcal{R} = \mathcal{D}(Y) \mapsto \mathbb{R} \rangle \tag{6}$$

To optimize performance in this setting, the agent searches for an optimal policy π^* that maximizes the expected reward under its own interventions:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(Y) \mid \mathcal{D}_{\text{exp}} \sim P_X(V) \right] \tag{7}$$

Unlike the off policy setting $\mathcal{T}_{\text{off}}$, online learning imposes no additional structural assumptions beyond the temporal order of states and actions. Therefore, the NUC condition is not required to hold. In fact, for any policy $\pi \in \Pi$, under the interventional model $\mathcal{M}_{\pi}^*$, all relevant variables S_i that influence each action X_i are assumed to be observable. As a result, the NUC condition is inherently satisfied in the post interventional setting. The following proposition formalizes this insight.

Lemma 1 (Experimental NUC)
Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where $\Pi = \{(X_i, S_i)\}_{X_i \in \boldsymbol{X}}$ and $\mathcal{R} : \mathcal{D}(Y) \to \mathbb{R}$. Then for any policy $\pi \in \Pi$, the NUC condition holds in the interventional model $\langle \mathcal{M}_{\pi}^*, \Pi, \mathcal{R} \rangle$ induced by $\texttt{do}(\pi)$.

3.2.1 Randomized Control trial (RCT)

We now describe algorithms that implement online policy learning through direct interaction with the environment. The first approach, *Randomized Controlled Trials (RCT)*, follows an "**explore then commit**" strategy. The agent explores the environment using randomly selected actions for a fixed number of episodes, then commits to the best performing policy based on data collected during exploration.

In each episode $t = 1, 2, \ldots$, the agent selects a policy $\pi^{(t)}$, performs an intervention via $\text{do}(\pi^{(t)})$, and observes the outcome $V^{(t)} \sim P_{\pi^{(t)}}(V)$.

The exploration phase lasts for N episodes. During this phase, the agent uses a *uniform random policy* π_{UNIF}, which assigns values to each action X_i by sampling uniformly from its domain. According to Lemma 1, the NUC condition holds under such interventions, enabling the agent to evaluate

any candidate policy $\pi \in \Pi$ using the IPW or DP methods described in Theorems 2 and 3.

After N episodes, the agent selects the policy with the highest empirical reward estimate and commits to it for the remainder of the episodes.

Algorithm 2: Randomized Controlled Trials (RCT)

Input: *Policy space* Π*, total number of exploration episodes* $N \in \mathbb{N}$

1. *For each episode* $t = 1, 2, \ldots$*:*

   (a) *Choose a policy* $\pi^{(t)}$ *as follows:*

   - *If* $t \leq N$*: use uniform random policy*

     $$\pi_{UNIF} = (X_1 \sim Unif(\mathcal{D}(X_1)), \ldots, X_H \sim Unif(\mathcal{D}(X_H)))$$

     and set $\pi^{(t)} \leftarrow \pi_{UNIF}$
   - *Else: set*

     $$\pi^{(t)} \leftarrow \arg\max_{\pi \in \Pi} \hat{\mathbb{E}}_{\pi}^{(N)}[Y]$$

   (b) *Perform intervention* $do(\pi^{(t)})$ *and receive observation* $V^{(t)} \sim P_{\pi^{(t)}}(V)$

3.2.2 The Upper Confidence Bound Algorithm

The *Upper Confidence Bound (UCB)* algorithm is built on the principle of *optimism in the face of uncertainty* (OFU), where the agent assumes the most favorable outcome consistent with past observations. Compared to RCT, UCB offers key advantages:

- It does not require prior knowledge of the reward gap between optimal and suboptimal policies.
- It operates without knowledge of the total number of episodes T.
- It achieves similar theoretical performance as RCT when tuned optimally.

UCB uses *adaptive randomization*, adjusting action probabilities based on prior outcomes. In a multi armed bandit (MAB) model $\mathcal{M}_{\text{MAB}} = (\mathbf{X}, \{P_x\}_{x \in \mathbf{X}}, Y)$, the agent computes an upper confidence bound for each arm x that overestimates the expected reward $\mathbb{E}_x[Y]$ with high probability.

Let $Y^{(1)}, \ldots, Y^{(n)}$ be i.i.d. samples from $P(Y)$, with empirical mean $\hat{\mathbb{E}}[Y] = \frac{1}{n}\sum_{i=1}^{n} Y^{(i)}$. Hoeffding's inequality implies:

$$\mathbf{P}\left(\mathbb{E}[Y] \leq \hat{\mathbb{E}}[Y] + \sqrt{\frac{\log(1/\delta)}{2n}}\right) \geq 1\delta, \quad \text{for } \delta \in (0,1)$$

Define $N_t(x) = \sum_{i=1}^{t} \mathbb{1}\{X^{(i)} = x\}$ as the count of times arm x was selected by time t. Then the UCB estimate at time t is:

$$\text{UCB}_t(x,\delta) = \hat{\mathbb{E}}_x^{(t)}[Y] + \sqrt{\frac{\log(1/\delta)}{2N_t(x)}}$$

where the empirical reward is given by:

$$\hat{\mathbb{E}}_x^{(t)}[Y] = \frac{1}{N_t(x)} \sum_{i=1}^{t} Y^{(i)} \cdot \mathbb{1}\{X^{(i)} = x\}$$

The first term estimates the reward; the second reflects the uncertainty (exploration bonus). At each episode, UCB selects the arm with the highest bound, performs do(x), and observes the reward $Y^{(t)}$.

Setting $\delta = t^4$ ensures increasingly tight bounds as learning progresses. This balances *exploration* (trying under sampled arms) with *exploitation* (choosing empirically high reward arms).

Algorithm 3: Upper Confidence Bound (UCB) in MAB

Input: *Action space* $\mathbf{X}$

1. *For each episode* $t = 1, 2, \ldots$ *:*
   (a) *Set confidence level* $\delta = t^4$
   (b) *For each arm* $x \in \mathbf{X}$*, compute:*

$$UCB_{t1}(x,\delta) = \hat{\mathbb{E}}_x^{(t1)}[Y] + \sqrt{\frac{\log(1/\delta)}{2N_{t1}(x)}}$$

   where $\hat{\mathbb{E}}_x^{(t1)}[Y]$ *is the empirical mean and* $N_{t1}(x)$ *the number of times* x *was selected up to* $t1$
   (c) *Select action:*

$$x^{(t)} \leftarrow \arg\max_{x \in \mathbf{X}} UCB_{t1}(x,\delta)$$

   (d) *Perform do*$(x^{(t)})$ *and observe reward* $Y^{(t)}$

3.3 Causal Identification

Online learning algorithms ensure the *No Unmeasured Confounding* (NUC) condition holds by deploying candidate policies in the environment. However, randomized experiments are not always feasible or ethical. For example, the causal effect of a risk factor such as smoking cannot be ethically investigated using randomized controlled trials. Moreover, such experiments can be prohibitively expensive. In a survey of phase trials for new drugs approved by the U.S. Food and Drug Administration (FDA) from 2015–2016, the median cost per patient was estimated at $ 41,117. These limitations highlight the need for policy learning approaches that do not rely on direct interventions.

An alternative approach is to relax the NUC assumption and leverage structural assumptions encoded in a causal diagram. This leads to the framework of *causal identification*, characterized by the signature:

$$\mathcal{T}_{\text{id}} = \langle \mathcal{L} = \texttt{see},\ \mathcal{A} = \mathcal{G},\ \Pi = \{(X_i, S_i)\}_{i=1}^{H},\ R = \mathcal{D}(Y) \to \mathbb{R} \rangle \tag{8}$$

The optimization objective is then:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}^{M_\pi}[R(Y)] \quad \text{subject to } A = G,\ D_{\text{obs}} \sim P(V) \tag{9}$$

Like in off policy learning , the agent collects observational data from the structural causal model (SCM) M over repeated episodes, observing samples from $P(V)$ without performing interventions ($I = \texttt{see}$). Crucially, the agent does *not* assume the NUC condition. Instead, it incorporates expert domain knowledge via a causal diagram G (see Def. 6 and Sec. 2.3 for discussion), assumed to be compatible with the true SCM.

The goal is to learn an optimal policy $\pi^* \in \Pi$ using both the observational data $P(V)$ and the structural assumptions in G. A central challenge is determining whether the interventional distribution $P_\pi(Y)$ can be uniquely recovered from $P(V)$ under the assumptions encoded in G, a property formalized by the notion of *identifiability*.

Definition 11 (Identifiability)
Let $X, Y \subseteq V$, and let π be a policy over X. We say the interventional distribution $P_\pi(Y)$ is identifiable from structural assumptions $\mathcal{A}$ if it is uniquely computable from any positive observational distribution $P(V)$ in any SCM M satisfying $\mathcal{A}$. That is, for all SCMs M_1, M_2 satisfying $\mathcal{A}$:

$$P_\pi(Y; M_1) = P_\pi(Y; M_2) \quad \textit{whenever} \quad P(V; M_1) = P(V; M_2) > 0.$$

In causal identification, the structural assumptions $\mathcal{A}$ are encoded in a causal graph G. We say $P_\pi(Y)$ is identifiable from G if the above condition

holds for all SCMs M_1, M_2 such that $G(M_1) = G(M_2) = G$. In contrast, off policy learning relies on the NUC condition as its structural assumption, restricting the form of the behavior policy f_X.

It follows, under the NUC condition, the expected reward $\mathbb{E}_\pi[R(Y)]$ is identifiable:

Corollary 1
Let $\Pi = \{(X_i, S_i)\}_{i=1}^H$ be a policy space and $R : D(Y) \to \mathbb{R}$ a reward function. Then, for any policy $\pi \in \Pi$, the interventional distribution $P_\pi(Y)$ is identifiable from the NUC condition (Def. 13) with respect to Π.

3.3.1 Sequential Backdoor for Policy Evaluation

As outlined in Section, both Inverse Propensity Weighting and Direct Policy Learning are foundational methods in off policy evaluation, widely used in causal inference and reinforcement learning. It is therefore critical to understand under which conditions these techniques can correctly identify expected rewards for a given policy space Π, when working with an arbitrary causal diagram $\mathcal{G}$.

To address this, we generalize a key graphical criterion known as the **sequential backdoor condition**, which helps determine whether the effects of executing a sequence of interventions can be identified using covariate adjustment. Unlike the original setting, which focuses on atomic policies, our version supports more general sequential policies, those that choose actions based on observed covariates.

Manipulated Graphs from Policy Interventions

Let $\pi \in \Pi$ be a policy. Denote by $\mathcal{G}_\pi$ the causal graph resulting from the intervention $\text{do}(\pi)$ in the underlying structural causal model. To construct $\mathcal{G}_\pi$ from $\mathcal{G}$:

1. For each timestep $i = 1, \ldots, H$, remove all incoming edges to the action variable X_i.

2. Add edges from the state variables S_i to the action node X_i as specified by the policy.

For a sub policy $\pi_{i:j} = (\pi_i, \ldots, \pi_j)$ with $1 \leq i < j \leq H$, we define the corresponding manipulated graph $\mathcal{G}_{\pi_{i:j}}$ as the diagram resulting from the intervention $\text{do}(\pi_i, \ldots, \pi_j)$.

3.3.2 Sequential Backdoor Condition

Definition 12 (Sequential Backdoor Condition)
Let $\mathcal{G}$ be a causal diagram and $Y \subseteq V$ denote the reward variables. A policy space $\Pi = \{(X_i, S_i)\}_{i=1}^{H}$ satisfies the sequential backdoor condition with respect to Y in $\mathcal{G}$ (also called backdoor admissible*) if for every policy $\pi \in \Pi$, and for all timesteps $i = 1, \ldots, H$, the following holds:*

$$X_i \perp\!\!\!\perp \boldsymbol{Y} \mid \overline{\boldsymbol{X}}_{i1}, \overline{\boldsymbol{S}}_i \quad \text{in } \mathcal{G}_{\underline{X_i}, \pi_{i+1:H}}$$

That is, conditioning on the history of past actions $\bar{X}_{i1}$ and state variables $\bar{S}_i$ blocks all backdoor paths from X_i to Y in the manipulated graph resulting from the remaining policy $\pi_{i+1:H}$.

This condition mirrors the classical backdoor criterion where a backdoor path is one that starts with an incoming edge into the variable of intervention (here, X_i). The sequential backdoor condition ensures that at each step, the observed history is sufficient to control for all confounding that may bias the estimate of X_i's effect on Y.

Special Case: Atomic Policies

Consider the case of an atomic policy $\pi = (\pi_1, \ldots, \pi_H)$ where each decision rule deterministically sets X_i to some fixed value x_i. For such policies, the manipulated graph $\mathcal{G}_{X_i, \ldots, X_H}$ is formed by simply removing all incoming edges to $X_{i+1}, \ldots, X_H$. In this case, the independence condition simplifies to:

$$X_i \perp\!\!\!\perp Y \mid \bar{X}_{i1}, \bar{S}_i \quad \text{in } \mathcal{G}_{\underline{X_i}, \overline{X_{i+1}, \ldots, X_H}}$$

This reduces to the original sequential backdoor criterion proposed by Pearl and Robins. Notably, the atomic version is a special case of the more general policy dependent version, implying that the generalized condition is strictly stronger.

Relation to the NUC Condition

If the No Unmeasured Confounding (NUC) condition holds, then all direct causes of each X_i are contained in the variables $\bar{X}_{i1}$ and $\bar{S}_i$, and there exist no unobserved confounders influencing both X_i and other variables. In this setting, conditioning on $\bar{X}_{i1}$ and $\bar{S}_i$ suffices to block all backdoor paths from X_i to Y.

Corollary 2. Let $\mathcal{M}$ be a CDM with associated causal diagram $\mathcal{G}$, and let the policy space $\Pi = \{(X_i, S_i)\}_{i=1}^{H}$ be backdoor admissible. If the NUC condition holds in $\mathcal{M}$, then Π automatically satisfies the sequential backdoor condition with respect to Y in $\mathcal{G}$.

Theorem 3
Let $\mathcal{G}$ be a causal diagram, and let $\Pi = \{\langle \mathbf{X}_i, \mathbf{S}_i \rangle\}_{i=1}^{H}$ denote a policy space. Suppose $R : \mathcal{D}(Y) \to \mathbb{R}$ is a reward function. If Π is backdoor admissible with respect to Y in $\mathcal{G}$ (Definition 12), then for any policy $\pi \in \Pi$, the expected reward $\mathbb{E}_\pi[R(Y)]$ is identifiable from $\mathcal{G}$. Furthermore, $\mathbb{E}_\pi[R(Y)]$ can be computed from the observational distribution $P(\mathbf{V})$ using either the IPW estimator or the DP estimator.

In essence, the sequential backdoor criterion extends conventional off policy evaluation methods to more general settings where the No Unmeasured Confounding (NUC) assumption is violated, and unobserved confounders may influence both actions and other variables in the system. Provided that the sequential backdoor condition is satisfied, one can employ IPW and DP estimators to assess candidate policies using observational data, while ensuring the asymptotic validity of the resulting estimates.

3.4 Novel Causal Reinforcement Learning Tasks

Recall that a CRL agent operates within a causal decision making model (CDM) denoted as $\langle \mathcal{M}, \Pi, \mathcal{R} \rangle$, where $\mathcal{M}$ represents an underlying structural causal model (SCM) that is not fully observable, Π is the policy space, and $\mathcal{R}$ is the reward function. Although the agent is ultimately evaluated in the true environment $\mathcal{M}$, for learning purposes, we replace $\mathcal{M}$ with a learning regime $\mathcal{L}$ and a set of structural assumptions $\mathcal{A}$ about the environment. This gives rise to a new task signature $(\mathcal{L}, \mathcal{A}, \Pi, R)$, which defines a causal reinforcement learning task.

The objective of the agent is to identify an optimal policy $\pi^\star$ within Π that maximizes the expected reward when evaluated under the true model $\mathcal{M}$, given assumptions $\mathcal{A}$ and access to data from $\mathcal{L}$:

$$\pi^\star = \arg\max_{\pi \in \Pi} \mathbb{E}_\pi^{\mathcal{M}} \left[R(Y) \mid \mathcal{A}, \mathcal{L} \right]$$

In essence, the CRL agent aims to optimize the policy with respect to the reward, while incorporating prior structural knowledge and leveraging data collected under a given learning regime.

A summary of the task signatures discussed so far is presented, which contextualizes the discussion. Each task examined represents a distinct point in the space defined by learning regimes and structural assumptions. For instance, off policy learning corresponds to a standard offline setting where the No Unmeasured Confounding (NUC) assumption is presumed valid. In such cases, classic estimators like IPW and DP can be used to generalize from observational data to new interventional policies.

We also explored online learning, where the data collection process is active and interventional. Here, fewer causal assumptions are needed since the data aligns directly with the intended inference. In contrast, causal identification represents a subtler offline setting where unconfoundedness might not hold, but sufficient structural knowledge allows the agent to either verify the NUC condition or optimize the policy objective regardless.

These three modalities, off policy learning, online learning, and causal identification, each embody different choices in the learning regime and structural assumptions about the true environment $\mathcal{M}$.

In the remainder of the paper, we introduce additional classes of CRL tasks that fall outside these traditional paradigms. These tasks are inspired by practical, real world settings and exhibit novel challenges and structural considerations, motivating new methods of analysis.

Novel CRL Task Dimensions

We now outline four novel causal reinforcement learning (CRL) problem settings that extend beyond standard learning paradigms. These tasks are motivated by practical limitations and challenges commonly faced in real world applications.

CRL 1: Causal Offline to Online Learning (COOL). [1] How can we pre train an online agent to accelerate its learning by using imperfect causal information derived from confounded observational data? Inferring the effects of candidate policies from observational data may be infeasible due to hidden confounding. Conversely, relying solely on trial and error experimentation is inefficient and potentially costly. The key challenge is to reduce the number of interventions required by the agent by leveraging invariant relationships encoded in a causal model. This task considers both online interaction (L2) and offline observational data (L1). We investigate methods that integrate these two modalities in settings where purely offline learning is provably inadequate, yet full scale online experimentation is undesirable.

CRL 2: Where to Intervene and What to Observe. [1] Should an agent intervene to achieve its goal? If so, where should the intervention be applied? The agent's objective is to discover an optimal policy from a set of candidates, each specifying actions and input states, including the option of inaction (null intervention). In the simplest case, the agent operates over a fixed action space Π_{EXP} and aims to identify an optimal intervention $\text{do}(X = x)$ that maximizes expected reward. We'll generalizes this to richer settings with a mixed action space Π_{MIX}, where interventions may differ qualitatively.

This scenario is particularly relevant in domains such as medicine, where evaluating combinations of treatments leads to combinatorial explosion in possible interactions.

CRL 3: Counterfactual Decision Making. [1] What if the agent had acted differently, would its outcome have improved? In this setting, the agent has executed an action $X = x$ and seeks to evaluate the counterfactual outcome under an alternative $\text{do}(X = x')$. Unlike prior settings where randomized experimental policies neutralize decision bias, we now consider learning under introspective regimes. We introduce the *Counterfactual Do* (Ctf Do) regime, where agents reason over a space of counterfactual policies Π_{CTF}. This capability becomes especially valuable in adversarial contexts, where an agent's tendencies are exploited, leading to systematic underperformance.

CRL 4: Causal Imitation Learning. [1] Does simply imitating expert behavior always yield effective policies? If not, under what conditions is imitation learning successful? Here, the agent must learn from demonstration data and a causal diagram when the reward function is ill defined or unavailable, and unmeasured confounding is present. In contrast to previous tasks that assumed known rewards, this setting reflects realistic challenges such as training autonomous systems from human demonstrations. For example, an autonomous vehicle may learn from a human driver's trajectories without explicit access to a reward metric. The challenge lies in enabling effective policy learning purely from observational traces and causal knowledge, without assuming access to a ground truth performance measure.

4 Causal Offline to Online Learning

The learning algorithms introduced in the previous section rely exclusively on a single mode of interaction with the environment, either through passive observation (i.e., "seeing" or offline data) or direct intervention (i.e., "doing" or online experimentation). Despite the strong theoretical guarantees of both approaches, a natural question arises: can an agent combine these two learning regimes to improve performance?

This question motivates the setting of *offline to online learning*. Existing approaches in the reinforcement learning literature assume the *No Unmeasured Confounding* (NUC) condition, making them unsuitable in scenarios where hidden confounders affect the data. In contrast, we study the problem of *causal offline to online learning* (COOL), where the agent begins with access to confounded observational data and then transitions to limited online experimentation. This setting relaxes the NUC assumption and reflects more realistic constraints encountered in practice.

Initially, the agent passively observes the environment over episodes $t = 1, \dots, n$, receiving observational samples $V^{(t)} \sim P(\mathbf{V})$. Beginning at episode $t = n + 1$, the agent selects a policy $\pi^{(t)}$ and interacts with the environment via direct intervention: $\text{do}(\pi^{(t)})$, observing outcomes $V^{(t)} \sim P_{\pi^{(t)}}(\mathbf{V})$. The agent aims to leverage the offline data $\{V^{(1)}, \dots, V^{(n)}\}$ to accelerate and guide the online learning process.

We formalize this hybrid learning setup with the following task signature:

$$\mathcal{T}_{\text{off+on}} = \langle \mathcal{L} = \{\texttt{see}, \texttt{do}\},\ \mathcal{A} = \emptyset,\ \Pi = \{(\mathbf{X}_i, \mathbf{S}_i)\}_{X_i \in \mathbf{X}},\ R : \mathcal{D}(Y) \to \mathbb{R} \rangle.$$

The objective of the agent is to identify a policy $\pi^\star$ that maximizes the expected reward, using both observational and experimental data:

$$\pi^\star = \arg\max_{\pi \in \Pi} \mathbb{E}_\pi^{\mathcal{M}} \left[R(Y) \mid \mathcal{D}_{\text{obs}} \sim P(\mathbf{V}),\ \mathcal{D}_{\text{exp}} \sim P_{\mathbf{x}}(\mathbf{V}) \right].$$

A distinguishing feature of this setting is the integration of both observational and interventional data to guide the learning process, enabling the agent to make more informed decisions while minimizing unnecessary interventions.

To formalize the intuition behind leveraging both observational and experimental data, we present an *online to offline* learning strategy that integrates observational data into the Upper Confidence Bound (UCB) algorithm, assuming that the No Unmeasured Confounding (NUC) condition holds.

Consider a multi armed bandit (MAB) model defined by $\langle \mathcal{M}, \{X, \emptyset\}, Y \rangle$. Let $\mathcal{D}_{\text{obs}} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{n}$ denote n i.i.d. samples drawn from the observational distribution $P(X, Y)$. At episode t, let $\mathcal{D}_{\text{exp}}^{(t)} = \{(X^{(n+i)}, Y^{(n+i)})\}_{i=1}^{t1}$ represent the experimental data collected by the UCB algorithm.

We define the empirical reward estimate for each arm x using both the observational and experimental data as follows:

$$\widehat{\mathbb{E}}_x^{(n+t)}[Y] = \frac{1}{N_{n+t}(x)} \left(\sum_{i=1}^{n} Y^{(i)} \cdot \mathbb{1}\{X^{(i)} = x\} + \sum_{i=1}^{t1} Y^{(n+i)} \cdot \mathbb{1}\{X^{(n+i)} = x\} \right),$$

Algorithm 4: Upper Confidence Bound in Direct Offline to Online Transfer (UCB)

Input: *Policy space* $\Pi = \{(X, \emptyset)\}$, *observational data* $\mathcal{D}_{obs} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{n}$

1: ***For*** *each episode* $t = 1, 2, \ldots$:

2: *Compute the augmented upper confidence bound for each arm* $x \in \mathbf{X}$:

$$UCB_{n+t}(x, \delta) = \widehat{\mathbb{E}}_x^{(n+t)}[Y] + \sqrt{\frac{\log(1/\delta)}{2N_{n+t}(x)}}$$

where $\delta = t^4$.

3: *Select the arm with the highest bound:*

$$X^{(n+t)} = \arg\max_{x \in \mathbf{X}} UCB_{n+t}(x, \delta)$$

4: *Perform the intervention* $do(X^{(n+t)})$ *and observe the reward* $Y^{(n+t)}$, *append to the data.*

5: ***End For***

This formulation allows the UCB algorithm to incorporate prior observational experience into its reward estimates while maintaining the exploration exploitation tradeoff during online learning. The integration is valid under the NUC assumption, which guarantees that the observational samples can be treated as if they were drawn from interventional distributions.

Let $N_{n+t}(x)$ denote the total number of times arm x appears in the com-

bined dataset of observational and experimental interactions:

$$N_{n+t}(x) = \sum_{i=1}^{n+t} \mathbb{1}\{X^{(i)} = x\},$$

where $\mathcal{D}_{\text{obs}} \cup \mathcal{D}_{\text{exp}}^{(t)}$ contains both observational samples and those collected online up to episode t.

Using this combined count, we define the *augmented upper confidence bound* for each arm x as:

$$\text{UCB}_{n+t}(x, \delta) = \widehat{\mathbb{E}}_x^{(n+t)}[Y] + \sqrt{\frac{\log(1/\delta)}{2N_{n+t}(x)}},$$

where $\widehat{\mathbb{E}}_x^{(n+t)}[Y]$ is the empirical reward estimate , and δ is a user defined confidence parameter.

This formulation defines an augmented version of the UCB algorithm, called UCB^+, that directly incorporates observational data as if it were collected under intervention, assuming the NUC condition holds. At each episode t, UCB^+ computes the augmented confidence bound $\text{UCB}_{n+t}(x, \delta)$ for each arm, selects the arm with the highest bound, and observes the resulting reward.

Performance Analysis. We now analyze the theoretical properties of UCB^+ under the NUC assumption. Let the MAB model be defined by $\langle \mathcal{M}, \{X, \emptyset\}, Y \rangle$, and assume that the NUC condition holds. The expected reward for each arm $x \in \mathbf{X}$ is identifiable from the observational distribution $P(X, Y)$ as:

$$\mathbb{E}_x[Y] = \mathbb{E}[Y \mid X = x].$$

Moreover, when evaluating an interventional policy $\pi(X)$ under the intervened model $(\mathcal{M}_\pi, \{X, \emptyset\}, Y)$, the NUC condition still holds, and the expected reward can also be computed from the interventional distribution:

$$\mathbb{E}_x[Y] = \mathbb{E}_\pi[Y \mid X = x].$$

These equalities justify pooling observational and interventional data to consistently estimate arm level expected rewards.

Equation above provides a consistent estimator for $\mathbb{E}_x[Y]$ under the NUC assumption. When enough observational data is available, UCB^+ can quickly identify the optimal arm with minimal online interaction. In such settings, off policy reward estimates derived from observational data can be used to initialize the online learning process, effectively *warm starting* the agent.

Robustness to Violations of NUC. Despite the advantages of this approach, the NUC assumption is often fragile and may not hold in many practical scenarios due to unmeasured confounding. Therefore, it is essential to empirically evaluate the performance of UCB * in settings where NUC is violated. We address this in the following experiments, which assess both the effectiveness and robustness of the augmented strategy under varying conditions.

In general, off policy estimation techniques may not reliably recover the unknown expected rewards of target policies in the absence of the NUC assumption. A straightforward transfer of estimated rewards can lead to inaccuracies in determining the optimal policy during the online learning phase, thereby degrading the performance of the learning algorithm. Furthermore, because the outcomes of interventions are not observed prior to the initiation of online learning, the learner is unable to identify potential biases introduced during the off policy evaluation stage using purely observational data. This presents substantial challenges for offline to online learning in settings where the NUC condition fails to hold. One might then conjecture that the learner should disregard all prior observational data, regardless of its volume, and initiate the online learning process anew.

This section demonstrates that such a conclusion is unwarranted and addresses the difficulties posed by the confounded scenario described above. We investigate the framework of *causal offline to online learning* (COOL), which seeks to enhance the efficiency of online reinforcement learning by incorporating observational offline data. Our focus lies in scenarios where the NUC assumption is violated, rendering the expected rewards of prospective policies non identifiable from observational data alone. In such cases, applying standard off policy evaluation directly can introduce substantial bias into reward estimates, ultimately impairing the subsequent online learning process.

4.1 Confounding Robust Offline to Online Learning

This section investigates the problem of offline to online learning in multi armed bandit (MAB) settings where the NUC assumption is violated, rendering conventional off policy learning methods, such as inverse propensity weighting (IPW, Theorem 2) and direct prediction (DP, Theorem 3), inapplicable. A natural question arises from the perspective of causal inference: can one estimate the expected rewards of actions from observational data using causal identification techniques, such as do calculus based approaches?

It is known, however, that in MAB environments, the expected rewards are generally not identifiable without introducing further assumptions. This

leads to the following corollary, which follows from the formal notion of identifiability presented in Definition 14.

Corollary 2 (Non Identifiability)
Consider endogenous variables $X, Y \subseteq V$ and let π be a policy over X. The interventional (policy) distribution $P_\pi(Y)$ is not identifiable from the structural assumptions $\mathcal{A}$ and the observational distribution $P(V)$ if there exist two structural causal models $\mathcal{M}_1, \mathcal{M}_2$ consistent with $\mathcal{A}$ such that

$$P(V; \mathcal{M}_1) = P(V; \mathcal{M}_2) > 0, \quad \textit{but} \quad P_\pi(Y; \mathcal{M}_1) \neq P_\pi(Y; \mathcal{M}_2).$$

Intuitively, this means that the expected reward $\mathbb{E}_x[Y]$ for an arm x cannot be identified if there exist two MAB environments that induce the same observational distribution $P(X, Y)$ but yield different expectations $\mathbb{E}_x[Y]$. Therefore, from observational data alone, the agent cannot uniquely recover the expected rewards of actions. The following example illustrates this non identifiability phenomenon in MAB settings.

The preceding discussion may suggest that in the presence of unobserved confounders and in the absence of additional causal assumptions, prior observational data cannot assist in evaluating the expected rewards of arms in MAB models. However, we demonstrate that this conclusion is overly pessimistic. It is possible to derive *causal bounds*, interval estimates on the unknown expected rewards, using observational data alone.

While the precise values of non identifiable expected rewards remain inaccessible, the learner can still infer partial information to constrain the feasible region of these expectations. Specifically, in MAB settings with action variable X and reward variable Y, foundational work by Manski (1990) and Robins (1989) provides techniques for deriving informative bounds, defined formally below, that contain the true expected reward $\mathbb{E}[Y \mid \text{do}(X = x)]$ using only the observational distribution. These causal bounds offer a principled way to leverage observational data in confounded environments.

Theorem 4 (Natural Bounds [2])
For any structural causal model $\mathcal{M}^*$ containing an action variable X and a reward variable Y, assume that the domain of X is discrete and finite, and that Y is bounded in the interval $[0, 1]$. Then, the expected reward for any arm x lies within the interval

$$\mathbb{E}_x[Y] \in [\ell_x, r_x],$$

where

$$\ell_x = \mathbb{E}[Y \mid x] \cdot P(x), \qquad r_x = \mathbb{E}[Y \mid x] \cdot P(x) + 1 - P(x),$$

and all terms are computed from the observational distribution.

The lower and upper bounds in Equation above depend solely on the observational distribution $P(X, Y)$ and are thus computable from observational data. These bounds are informative and lie strictly within the interval $[0, 1]$ whenever the marginal probabilities $P(x) > 0$ for all arms $x \in \mathcal{D}(X)$. Furthermore, the natural bounds have been shown to be *optimal* in the context of MAB models, meaning they cannot be tightened without introducing additional assumptions or access to interventional data.

The causal bounds $\mathbb{E}_x[Y] \in [\ell_x, r_x]$ can be leveraged to refine the upper confidence bound estimates assigned to each arm x during the online learning phase. Specifically, for a UCB algorithm selecting an arm at episode t, the causally clipped upper confidence bound for arm x is defined as

$$\overline{\text{UCB}}_t(x, \delta) = \min\left\{\max\left\{\text{UCB}_t(x, \delta), \ell_x\right\}, r_x\right\}.$$

In this formulation, $\text{UCB}_t(x, \delta)$ denotes the standard upper confidence bound used in MAB models, as introduced in Equation (182). The clipping ensures that the adjusted confidence bound $\overline{\text{UCB}}_t(x, \delta)$ remains within the causal interval $[\ell_x, r_x]$. Intuitively, whenever the confidence bound estimated from experimental data is inconsistent with the causal bounds, particularly in the early stages of learning when t is small and $\text{UCB}_t(x, \delta)$ is relatively loose, the causal bounds take precedence. As more experimental samples $N_t(x)$ are gathered over time, the standard upper bound progressively aligns with the causal interval.

The algorithm prioritizes the causal bounds because observational data is typically more abundant, whereas experimental interaction (interventions) can be costly. Since the bounds in Theorem 8 are valid and estimable from sufficient observational data, they offer a reliable means of constraining uncertainty during learning.

Algorithm 5 presents the augmented UCB algorithm that integrates causal bounds derived from observational data, referred to as UCB^+. It takes as input the causal bounds $\mathbb{E}_x[Y] \in [\ell_x, r_x]$ for all candidate arms x. At each episode t, it computes the clipped confidence bound $\overline{\text{UCB}}_{n+t}(x, \delta)$ by incorporating both the observationally derived causal bounds and experimental data accumulated up to that point. The algorithm then selects the arm with the highest clipped confidence bound and observes the resulting reward.

Algorithm 5: Causal-UCB for Multi-Armed Bandits

Input: *Action space $\mathcal{X}$; known causal bounds $\mathbb{E}_x[Y] \in [\ell_x, r_x]$ for each $x \in \mathcal{X}$*
For each round $t = 1, 2, \ldots$:

- *For each* $x \in \mathcal{X}$*, compute:*

$$\mathrm{UCB}_t(x) = \min\left\{r_x, \hat{\mu}_t(x) + \beta_t(x)\right\}$$

- *Choose arm:*

$$X(t) = \arg\max_{x \in \mathcal{X}} \mathrm{UCB}_t(x)$$

- *Perform intervention* $\mathrm{do}(X(t))$ *and observe reward* $Y(t)$
- *Update* $\hat{\mu}_{t+1}(X(t))$ *and* $\beta_{t+1}(X(t))$

4.2 Online Learning in Sequential Decision Making

The offline to online learning framework discussed thus far has focused on MAB models with a single decision point, i.e., a horizon of $H = 1$. In what follows, we generalize this framework to optimize a causal decision model (CDM) denoted by $(\mathcal{M}, \Pi, R)$, where the policy space is given by $\Pi = \{\pi_i(X_i \mid S_i)\}_{i=1}^{H}$, and the reward function is $R : \mathcal{D}(Y) \to \mathbb{R}$, mapping a set of signals Y to a scalar reward.

We begin by introducing a purely online learning algorithm designed to optimize a CDM $(\mathcal{M}, \Pi, R)$ without requiring explicit parametrization of the underlying environment $\mathcal{M}$. Unlike the previously discussed algorithms for MAB settings, our proposed method engages with the underlying structural causal model $\mathcal{M}$ over multiple episodes indexed by $t = 1, \ldots, T$.

In each episode t, instead of selecting a single action X, the algorithm determines a sequence of actions $(X_1, \ldots, X_H)$. More precisely, it chooses a policy $\pi^{(t)} = (\pi_1^{(t)}, \ldots, \pi_H^{(t)})$ at the start of episode t. For every decision stage $i = 1, \ldots, H$, the algorithm observes state $S_i^{(t)}$, selects an action $X_i^{(t)} \sim \pi_i^{(t)}(X_i \mid S_i^{(t)})$, and performs the intervention $\text{do}(X_i \leftarrow X_i^{(t)})$.

We now present a novel online learning algorithm designed to optimize a policy space over a sequence of actions $X = \{X_1, \ldots, X_H\}$.

We begin by introducing the necessary notation and technical preliminaries. For each stage $i = 1, \ldots, H$, define the action history up to stage i as $\bar{X}_i = \{X_1, \ldots, X_i\}$ and the corresponding state history as $\bar{S}_i = \{S_1, \ldots, S_i\}$.

For any policy $\pi \in \Pi$, by applying standard probabilistic identities and Bayes' rule, the expected reward can be expressed as

$$\mathbb{E}_\pi\left[\mathcal{R}(Y)\right] = \sum_{\bar{x}_H, \bar{s}_H} \underbrace{\mathbb{E}_\pi\left[\mathcal{R}(Y) \mid \bar{x}_H, \bar{s}_H\right]}_{\text{reward}} \prod_{i=0}^{H1} \underbrace{P_\pi\left(s_{i+1} \mid \bar{x}_i, \bar{s}_i\right)}_{\text{transition probabilities}} \underbrace{\pi_{i+1}\left(x_{i+1} \mid s_{i+1}\right)}_{\text{policy}}. \tag{10}$$

The above decomposition holds in the post interventional submodel $\mathcal{M}_\pi$, where each action X_i is determined by the policy function π_i. Among the quantities involved, the policy π is assumed known, and thus it suffices to estimate the transition distributions $P_\pi(s_{i+1} \mid \bar{x}_i, \bar{s}_i)$ and the conditional reward $\mathbb{E}_\pi[\mathcal{R}(Y) \mid \bar{x}_H, \bar{s}_H]$.

By Lemma 1, the NUC condition is satisfied in the post interventional system $(\mathcal{M}_\pi, \Pi, \mathcal{R})$. Consequently, for each $i = 0, \ldots, H1$, conditioning on the past states $\bar{S}_i$ d separates all backdoor paths between the actions $\bar{X}_i$ and any other variables in the submodel $\mathcal{M}^\pi$. By applying Rule 2 of the do calculus,

$$P_\pi(s_{i+1} \mid \bar{x}_i, \bar{s}_i) = P_{\bar{x}_i}(s_{i+1} \mid \bar{s}_i), \tag{11}$$
$$\mathbb{E}_\pi[\mathcal{R}(Y) \mid \bar{x}_H, \bar{s}_H] = \mathbb{E}_{\bar{x}_H}[\mathcal{R}(Y) \mid \bar{s}_H]. \tag{12}$$

In words, the transition distribution $P_\pi(s_{i+1} \mid \bar{x}_i, \bar{s}_i)$ and the conditional expected reward $\mathbb{E}_\pi[\mathcal{R}(Y) \mid \bar{x}_H, \bar{s}_H]$ are invariant across all policies $\pi \in \Pi$. This invariance implies that both quantities can be consistently estimated by pooling interventional data collected under different candidate policies in the policy space Π.

Algorithm 6: Upper Confidence Bound (UCB) for CDMs

Input: *A policy space* $\Pi = \{(X_i, S_i)\}_{i=1}^H$*, a reward function* $R : \mathcal{D}(Y) \to [0, 1]$.

1. *For each episode* $t = 1, 2, \ldots$*:*
   (a) *Construct the set* $\mathcal{M}_{t1}(\delta)$ *of all candidate SCMs* $\mathcal{M}$*, with error probability* $\delta = \frac{1}{t^4}$*, that are compatible with the interventional data* $\{V^{(1)}, \ldots, V^{(t1)}\}$ *collected up to episode* t*.*
   (b) *Find the optimal policy* $\pi^{(t)}$ *for an optimistic SCM* $\mathcal{M}^{(t)} \in \mathcal{M}_{t1}(\delta)$ *such that:*

$$\mathbb{E}_{\pi^{(t)}}[R(Y); \mathcal{M}^{(t)}] = \max_{\pi \in \Pi,\, \mathcal{M} \in \mathcal{M}_{t1}(\delta)} \mathbb{E}_\pi[R(Y); \mathcal{M}]$$

   (c) *Execute the policy* $\pi^{(t)}$ *via* $\texttt{do}(\pi^{(t)})$ *and observe* $V^{(t)}$*.*

Throughout this section, we assume that the state and action spaces, $\mathbf{S}$ and $\mathbf{X}$ respectively, are finite, and the reward function $R(Y)$ is bounded within the interval $[0, 1]$. Let $\pi^{(1)}, \ldots, \pi^{(t)} \in \Pi$ be a fixed sequence of policies. Given a finite set of samples $\{V^{(1)}, \ldots, V^{(t)}\}$ drawn from the corresponding interventional distributions $P_{\pi^{(1)}}(V), \ldots, P_{\pi^{(t)}}(V)$, we define the empirical estimators for the transition probabilities $P_{x_{1:i1}}(S_i \mid s_{1:i1})$, for $i = 1, \ldots, H$, and the conditional reward $\mathbb{E}_{x_{1:H}}[Y \mid s_{1:H}]$ as follows:

$$\forall i = 1, \ldots, H1, \quad \hat{P}^{(t)}_{\bar{x}_i}(s_{i+1} \mid \bar{s}_i) = \frac{\sum_{j=1}^t \mathbb{1}\left\{\bar{X}_i^{(j)} = \bar{x}_i,\ \bar{S}_{i+1}^{(j)} = \bar{s}_{i+1}\right\}}{N_t(\bar{x}_i, \bar{s}_i)},$$

$$\text{and} \quad \hat{\mathbb{E}}^{(t)}_{\bar{x}_H}[R(Y) \mid \bar{s}_H] = \frac{\sum_{j=1}^t \mathbb{1}\left\{\bar{X}_H^{(j)} = \bar{x}_H,\ \bar{S}_H^{(j)} = \bar{s}_H\right\} R(Y^{(j)})}{N_t(\bar{x}_H, \bar{s}_H)},$$

Algorithm above outlines a UCB based approach for optimizing an unknown causal decision model (CDM) Π, action space Π, and reward function R. The algorithm operates in phases: model construction, optimistic planning, and policy execution.

In Step 2, the algorithm builds a set $\mathcal{M}_{t1}(\delta)$ of plausible structural causal models (SCMs) using interventional data $\{V^{(1)}, \ldots, V^{(t1)}\}$. For each SCM $M \in \mathcal{M}_{t1}(\delta)$, the transition probabilities $P_{\bar{x}_i}(S_{i+1} \mid \bar{s}_i)$ for $i = 1, \ldots, H+1$, and the conditional reward $\mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H]$, lie within convex intervals centered around their empirical counterparts derived from previously collected interventional samples.

The confidence parameter δ is designed to decrease with the episode index t, so that $\mathcal{M}_{t1}(\delta)$ captures the true SCM Π with high probability as more data is gathered. In Step 3, the most optimistic SCM $M^{(t)} \in \mathcal{M}_{t1}(\delta)$ is selected, and its optimal policy $\pi^{(t)}$ is computed to maximize expected reward. Step 4 then executes this policy throughout episode t to collect new interventional data $V^{(t)} \sim P^{\pi^{(t)}}(V)$.

The discrepancy between the true and empirical distributions is bounded in L_1 norm using a concentration inequality from Weissman et al. (2003):

$$\mathbb{P}\left(\|\hat{P}(\cdot)P(\cdot)\|_1 > \sqrt{\frac{2|\mathcal{D}(Y)|\log(2/\delta)}{n}} \right) \leq \delta$$

Fix a confidence level $\delta \in (0,1)$. Define $\mathcal{M}_t(\delta)$ as the set of SCMs M with endogenous variables V, such that for each $i = 0, \ldots, H+1$, the transition distribution $P_{\bar{x}_i}(S_{i+1} \mid \bar{s}_i)$ is close to the empirical estimate $\hat{P}^{(t)}_{\bar{x}_i}(S_{i+1} \mid \bar{s}_i)$, and the expected reward $\mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H]$ is close to its empirical estimate $\hat{\mathbb{E}}^{(t)}_{\bar{x}_H}[R(Y) \mid \bar{s}_H]$. Specifically:

$$\|P_{\bar{x}_i}(\cdot \mid \bar{s}_i; M)\hat{P}^{(t)}_{\bar{x}_i}(\cdot \mid \bar{s}_i)\|_1 \leq \varsigma_i(\delta)$$

$$\left| \mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H; M]\hat{\mathbb{E}}^{(t)}_{\bar{x}_H}[R(Y) \mid \bar{s}_H] \right| \leq \varsigma_H(\delta)$$

The confidence radius $\varsigma_i(\delta)$ is defined for $i = 0, \ldots, H$ as:

$$\varsigma_i(\delta) = L\sqrt{\frac{2|\mathcal{D}(S_{i+1})|\log\left(\frac{4H|\mathcal{D}(\bar{S}_i \to \bar{X}_i)|}{\delta}\right)}{N_t(\bar{x}_i, \bar{s}_i)}}$$

and for rewards:

$$\varsigma_H(\delta) = L\sqrt{\frac{\log\left(\frac{8H|\mathcal{D}(\bar{S}_H \to \bar{X}_H)|}{\delta}\right)}{2N_t(\bar{x}_H, \bar{s}_H)}}$$

By applying a union bound over these concentration bounds and Hoeffding's inequality, the probability that the true model Π is excluded from $\mathcal{M}_t(\delta)$ is at most:

$$\mathbb{P}(\Pi \notin \mathcal{M}_t(\delta)) \leq \sum_{i=0}^{H} \sum_{\bar{x}_i, \bar{s}_i} \frac{\delta}{2H|\mathcal{D}(\bar{S}_i \to \bar{X}_i)|} + \sum_{\bar{x}_H, \bar{s}_H} \frac{\delta}{2H|\mathcal{D}(\bar{S}_H \to \bar{X}_H)|} < \delta$$

Thus, the true SCM Π belongs to the candidate model set $\mathcal{M}_t(\delta)$ with probability at least 1δ.

The UCB algorithm performs optimistic planning by identifying a policy $\pi^{(t)}$ that is optimal for a carefully chosen optimistic SCM $M^{(t)}$. Given a fixed SCM M, standard planning techniques (e.g., dynamic programming from Bellman, 1957; Koller and Milch, 2003) can be employed to determine a policy over the action space Π that maximizes the expected reward.

However, the challenge here extends beyond policy optimization for a fixed model. It also involves selecting the particular SCM $M^{(t)}$ from the set of plausible models $\mathcal{M}_{t1}(\delta)$ such that the resulting optimal policy achieves the highest possible expected reward.

This dual layer optimization problem—over both models and policies—can be cast as a polynomial programming problem. Leveraging the decomposition structure and model invariances, the expected reward under a policy π can be expressed as:

$$\mathbb{E}_\pi[R(Y)] = \sum_{\bar{x}_H, \bar{s}_H} \mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H] \prod_{i=0}^{H} P_{\bar{x}_i}(s_{i+1} \mid \bar{s}_i) \pi_{i+1}(x_{i+1} \mid s_{i+1})$$

This representation shows that the expected reward decomposes into components involving transition probabilities and the policy itself. The learner therefore seeks the SCM $M^{(t)}$ and corresponding policy $\pi^{(t)}$ that jointly maximize this quantity over the feasible set $\mathcal{M}_{t1}(\delta)$.

Each decision rule $\pi_i(X_i \mid S_i)$ for $i = 1, \ldots, H$ is defined as a proper conditional probability distribution that maps states S_i to actions X_i. The transition probabilities $P_{\bar{x}_i}(S_{i+1} \mid \bar{s}_i)$, for $i = 0, \ldots, H1$, lie within the convex confidence set characterized earlier, while the expected rewards $\mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H]$ lie within a convex polytope $\mathcal{R}$ defined previously.

Solving for both an optimistic SCM $M^{(t)}$ and a corresponding policy $\pi^{(t)}$ is equivalent to solving a polynomial program. The objective of this program is to maximize the expected reward $\mathbb{E}_\pi[R(Y)]$, where the transition probabilities $P_{\bar{x}_i}(S_{i+1} \mid \bar{s}_i)$ and expected reward $\mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H]$ are denoted by variables Π_i and R, respectively. The program is subject to standard

probabilistic constraints:

$$\sum_{x_i} \pi_i(x_i \mid s_i) = 1, \quad \pi_i(x_i \mid s_i) \geq 0, \quad \text{for all } i = 1, \ldots, H.$$

When the policy space Π satisfies the perfect recall condition—i.e., the agent retains access to all prior observations and decisions—the polynomial program can be solved efficiently using dynamic programming techniques (Koller and Friedman, 2009, Def. 23.5). This condition implies that $S_i \subseteq \{X_1, \ldots, X_{i1}\} \cup \{S_1, \ldots, S_{i1}\}$ for $i < j$, meaning no past information is forgotten.

An optimistic policy $\pi^{(t)}$ can then be computed by solving the following extended Bellman style recursive equations for $i = 1, \ldots, H1$:

$$Q(s_i, \bar{x}_{i1}) = \max_{x_i} \left\{ \max_{P_{\bar{x}_i}(\cdot|\bar{s}_i) \in \Pi_i} \left[\sum_{s_{i+1}} Q(s_{i+1}, \bar{x}_i) \cdot P_{\bar{x}_i}(s_{i+1} \mid \bar{s}_i) \right] \right\},$$

and for the terminal step $i = H$:

$$Q(s_H, \bar{x}_{H1}) = \max_{x_H} \left\{ \max_{\mathbb{E}_{\bar{x}_H}[R(Y)|\bar{s}_H] \in \mathcal{R}} \mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H] \right\}.$$

4.3 Learning from Observational Data

Despite offering strong performance guarantees, the online learning algorithm presented in Algorithm above does not utilize information from the observational distribution $P(V)$. Under the *No Unmeasured Confounding* (NUC) condition in the underlying Causal Decision Model (CDM) $(M^\star, \Pi, R)$, the history of states and actions $(\bar{X}_{i1}, \bar{S}_i)$ blocks all backdoor paths from each action X_i to any other variable in the causal graph $\mathcal{G}$. Consequently, it becomes valid to estimate the interventional transition distribution $P_{\bar{x}_i}(S_{i+1} \mid \bar{s}_i)$ for $i = 0, \ldots, H1$, and the expected reward $\mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H]$, using their corresponding observational conditional distributions:

$$P(S_{i+1} \mid \bar{x}_i, \bar{s}_i) \quad \text{and} \quad \mathbb{E}[R(Y) \mid \bar{x}_H, \bar{s}_H].$$

The validity of this substitution follows directly from Rule 2 of do calculus. These observational estimates can serve as a warm start for the UCB algorithm, allowing it to begin learning with some prior knowledge.

However, in general scenarios where the NUC condition is violated and no additional causal assumptions are available, these quantities may not be identifiable. This leads to challenges in estimation due to potential confounding.

To address such cases, we turn to the notion of *partial identification*. Specifically, we aim to bound the interventional transition probabilities $P_{\bar{x}_i}(s_{i+1} \mid \bar{s}_i)$ and expected rewards from the observational distribution $P(V)$, even in the presence of unmeasured confounding.

Theorem 5
Let $(M^\star, \Pi, R)$ be a Causal Decision Model (CDM), where the policy space is

$$\Pi = \{\pi_i(X_i \mid S_i)\}_{i=1}^{H}.$$

For every $i = 1, \ldots, H1$, the interventional transition probability

$$P_{\bar{x}_i}(s_{i+1} \mid \bar{s}_i)$$

is bounded as

$$P_{\bar{x}_i}(s_{i+1} \mid \bar{s}_i) \in [l_{\bar{x}_i}(s_{i+1}),\ r_{\bar{x}_i}(s_{i+1})],$$

where

$$l_{\bar{x}_i}(s_{i+1}) = \frac{\Gamma(\bar{s}_{i+1}, \bar{x}_i)}{\Gamma(\bar{s}_i, \bar{x}_{i1})}, \quad r_{\bar{x}_i}(s_{i+1}) = \frac{\Gamma(\bar{s}_{i+1}, \bar{x}_i)}{\Gamma(\bar{s}_i, \bar{x}_{i1})}.$$

The function Γ is defined recursively using the observational distribution $P(V)$ as:

$$\Gamma(\bar{s}_{i+1}, \bar{x}_i) = \begin{cases} P(s_1), & \text{if } i = 0, \\ P(\bar{s}_{i+1}, \bar{x}_i)P(\bar{s}_i, \bar{x}_i) + \Gamma(\bar{s}_i, \bar{x}_{i1}), & \text{if } i \geq 1. \end{cases}$$

The upper bound in Eq. (283) can be rewritten as:

$$\frac{\Gamma(\bar{s}_{i+1}, \bar{x}_i)}{\Gamma(\bar{s}_i, \bar{x}_{i1})} = \frac{P(\bar{s}_{i+1}, \bar{x}_i)P(\bar{s}_i, \bar{x}_i) + \Gamma(\bar{s}_i, \bar{x}_{i1})}{\Gamma(\bar{s}_i, \bar{x}_{i1})}$$

$$= 1 + \frac{P(\bar{s}_{i+1}, \bar{x}_i)P(\bar{s}_i, \bar{x}_i)}{\Gamma(\bar{s}_i, \bar{x}_{i1})}$$

Interpretation. Consider the denominator. Observe that the gap $P(\bar{s}_i, \bar{x}_i) \Rightarrow P(\bar{s}_{i+1}, \bar{x}_i) > 0$ whenever the observational probabilities $P(s, x)$ are strictly positive for all realizations of the state action pair. This implies that the bounds provided in above Theorem are generally informative, i.e., they are strictly contained in the interval $[0, 1]$.

The bounds presented in Theorem above can thus be seen as a generalization of the natural causal bounds from Natural Bounds, now extended to sequential decision settings where the number of actions $H > 1$.

Theorem 6
Let $\mathcal{M} = (\mathcal{G}, \Pi, \mathcal{R})$ be a Causal Decision Model (CDM), where the policy space $\Pi = \{\pi_i(X_i \mid S_i)\}_{i=1}^{H}$ and the reward function $R : \mathcal{D}(Y) \to [0, 1]$. Then the interventional reward mean $\mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H]$ is bounded as:

$$\mathbb{E}_{\bar{x}_H}[R(Y) \mid \bar{s}_H] \in [\ell_{\bar{x}_H}(\bar{s}_H),\ r_{\bar{x}_H}(\bar{s}_H)]\,,$$

where

$$\ell_{\bar{x}_H}(\bar{s}_H) = \mathbb{E}[R(Y) \mid \bar{s}_H, \bar{x}_H] \cdot \frac{P(\bar{s}_H, \bar{x}_H)}{\Gamma(\bar{s}_H, \bar{x}_H)},$$

$$r_{\bar{x}_H}(\bar{s}_H) = (\mathbb{E}[R(Y) \mid \bar{s}_H, \bar{x}_H]) \cdot \frac{P(\bar{s}_H, \bar{x}_H)}{\Gamma(\bar{s}_H, \bar{x}_{H \Rightarrow 1})} + \mathbb{E}[R(Y) \mid \bar{s}_H, \bar{x}_H].$$

Since the conditional reward $\mathbb{E}[R(Y) \mid \bar{s}_H, \bar{x}_H] \in [0, 1]$, the bounds presented must lie strictly within the interval $[0, 1]$ whenever the observational probability $P(\bar{s}_H, \bar{x}_H)$ is non zero. This ensures that the bounds are informative. These bounds are expressed as functions of the observational distribution $P(V)$, which is identifiable through the data collection process and can, therefore, be consistently estimated. Consequently, the causal bounds can be approximated using the corresponding sample mean estimates. Furthermore, standard concentration inequalities can be employed to quantify and control the uncertainty arising from finite sample sizes.

Algorithm 7: Upper Confidence Bound with Causal Bounds (UCB$^+$)

Require: *a policy space* $\Pi = \{(X_i, S_i)\}_{i=1}^{H}$, *a reward function* $\mathcal{R} : \mathcal{G}(\mathcal{Y}) \mapsto [0, 1]$, *causal bounds* $[\underline{\ell}_x(\bar{s}_{i+1}), \bar{r}_x(\bar{s}_{i+1})]$, *for* $i = 1, \ldots, H1$, *and* $[\underline{\ell}_{x_H}(\bar{s}_H), \bar{r}_{x_H}(\bar{s}_H)]$ *for all state action pairs* $(x, s) \in \mathcal{X} \times \mathcal{S}$.

1. *Let* $\mathcal{M}_c$ *be the collection of all SCMs* $\mathcal{M}$ *with endogenous variables* $\mathbf{V}$, *and transition distributions* $P_{x_i}(S_{i+1} \mid \bar{s}_i; \mathcal{M})$ *consistent with bounds* $[\underline{\ell}_x(\bar{s}_{i+1}), \bar{r}_x(\bar{s}_{i+1})]$ *for* $i = 0, \ldots, H1$, *and reward expectations* $\mathbb{E}_{x_H}[Y \mid \bar{s}_H; \mathcal{M}]$ *consistent with* $[\underline{\ell}_{x_H}(\bar{s}_H), \bar{r}_{x_H}(\bar{s}_H)]$, *such that:*

$$\forall i = 0, \ldots, H1, \quad \underline{\ell}_x(\bar{s}_{i+1}) \leq P_{x_i}(S_{i+1} \mid \bar{s}_i; \mathcal{M}) \leq \bar{r}_x(\bar{s}_{i+1}), \tag{13}$$

$$\underline{\ell}_{x_H}(\bar{s}_H) \leq \mathbb{E}_{x_H}[Y \mid \bar{s}_H; \mathcal{M}] \leq \bar{r}_{x_H}(\bar{s}_H). \tag{14}$$

2. ***for*** *each episode* $t = 1, 2, \ldots$ ***do***

3. *Construct a set of plausible SCMs* $\mathcal{M}_{t1}(\delta)$ *using* $\delta = t^4$, *as in Steps 2 4 of Algorithm 6.*

4. *Identify the optimal policy* $\pi^{(t)}$ *of an optimistic SCM* $\mathcal{M}^{(t)} \in \mathcal{M}_{t1}(\delta) \cap \mathcal{M}_c$ *such that:*

$$\mathbb{E}_{\pi^{(t)}, \mathcal{M}^{(t)}}[Y; \mathcal{M}^{(t)}] = \max_{\pi, \mathcal{M}} \mathbb{E}_{\pi, \mathcal{M}}[Y; \mathcal{M}] \quad s.t.\ \pi \in \Pi, \mathcal{M} \in \mathcal{M}_{t1}(\delta) \cap \mathcal{M}_c. \tag{15}$$

5. *Execute* $\pi^{(t)}$ *for episode* t *and collect observations* $V^{(t)}$.

6. ***end for***

5 Mixed Policy Learning: When and Where to Intervene

Agents operating in complex, uncertain environments must process large volumes of information efficiently and safely. Achieving this requires identifying optimal policies that drive desirable outcomes. Most existing work, including previous sections, assumes a fixed action space—often over a variable set X with $|X| = k$ limiting policies to 2^k configurations in binary settings.

This section relaxes the fixed action assumption by considering flexible action spaces, including settings where interventions are optional. While full intervention may be feasible in controlled environments (e.g., simulators), it can be risky or impractical in real world systems due to potential side effects.

Interestingly, in non Markovian causal systems, full control via $\text{do}(X \leftarrow x)$ is not always optimal. In some cases, partial interventions—targeting a subset $X^{\text{`}} \subseteq X$ can yield better outcomes. However, this flexibility increases complexity, requiring evaluation over up to 3^k intervention subsets.

Real world dynamics often reduce the need for full control. For example, physical constraints (e.g., inertia, gravity) naturally influence robotic behavior, and medical decisions depend on practitioner expertise and regulatory frameworks. In such scenarios, intervening on a subset of variables while allowing others to evolve naturally under their inherent mechanisms f_X can be more effective.

To formalize this setting, we introduce a *mixed policy space*, which comprises multiple policy spaces defined over subsets of action and context variables:

$$\Pi_{\text{MIX}} = \left\{ \{\pi_{X^{\text{`}}, S_{X^{\text{`}}}}\} \mid X^{\text{`}} \subseteq X_\star, S_{X^{\text{`}}} \subseteq S_\star \right\}. \tag{16}$$

Each element in Π_{MIX} spans a spectrum from purely observational to fully experimental policies.

For convenience, we use the shorthand $\pi \in \Pi_{\text{MIX}}$ to denote $\pi \in \Pi \subseteq \Pi_{\text{MIX}}$.

The mixed policy learning problem is characterized by the task signature:

$$\mathcal{T}_{\text{MIX}} = (\mathcal{I} = \texttt{do},\ \mathcal{A} = \mathcal{G},\ \Pi_{\text{MIX}} = \{\{\pi_{X_i, S_i} \mid S_i \subseteq S_\star\}_{X_i \subseteq X}\},\ X \subseteq X_\star,\ R).$$

Unlike standard settings with a fixed policy space, here the agent searches over a mixed policy space to find:

$$\pi^\star = \arg\max_{\pi \in \Pi_{\text{MIX}}} \mathbb{E}_\pi^{\mathcal{M}}\left[R(Y)\right],$$

where the expectation is evaluated under the interventional distribution induced by the observational data $D^{\text{exp}} \sim P_x(V)$, and the causal graph $\mathcal{G}$. This task is distinguished by its flexible policy scope.

5.1 Mixed Policy with no context

We now explore how an agent can efficiently discover the optimal action in environments defined by a mixed policy space and an underlying causal diagram. For simplicity, we consider a multi armed bandit (MAB) setting where each policy $\pi \in \Pi_{\text{MIX}}$ is an experimental policy over a subset $X^{\text{`}} \subseteq X^{\star}$, with each action rule $\pi(x_{\text{`}})$ defined without contextual information.

Example 1 (Where to Intervene)
Consider a MAB environment governed by the structural causal model $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$*, where:*

$$\mathbf{U} = \{U_1, U_2\}, \quad \mathbf{V} = \{X_1, X_2, Y\},$$

and all variables are binary. The causal mechanisms are defined as:

$$F = \begin{cases} X_1 \leftarrow U_1, \\ X_2 \leftarrow X_1 \oplus U_2, \\ Y \leftarrow X_2 \oplus U_2, \end{cases}$$

with independent noise: $P(U_1) = P(U_2) = \frac{1}{2}$.

An agent deployed in this environment aims to maximize Y *by choosing optimal interventions over* X_1 *and* X_2*:*

$$\pi^{\star} = \arg\max_{x_1, x_2} \mathbb{E}_{do(X_1 = x_1, X_2 = x_2)}[Y; \mathcal{M}].$$

The mixed policy space reflects all subsets of the action set $X_{\vartheta} = \{X_1, X_2\}$*, resulting in* $2^{|X_{\vartheta}|}$ *possible intervention scopes. This maps naturally to a MAB formulation, where each arm corresponds to an intervention set and assignment (e.g.,* $do(X_1 = 0), do(X_2 = 1), \ldots$*), leading to 9 distinct arms.*

We analyze a naive strategy in which the agent lacks knowledge of the causal graph and follows a strict experimentalist approach, intervening simultaneously on all action variables $X = \{X_1, X_2\}$*. We refer to this as the* all at once *agent. Its actions correspond to:*

$$do(X_1 = x_1, X_2 = x_2), \quad \textit{for all } x_1, x_2 \in \{0, 1\}.$$

While the agent's internal model assumes a flat structure (see Fig. 29b), its performance is still governed by the actual SCM $\mathcal{M}$*. A key question arises: Can ignorance of the causal structure be compensated by conducting more interventions? That is, can repeated sampling from* $do(X_1 = x_1, X_2 = x_2)$ *distributions lead to the correct policy?*

After interacting with the environment, the all at once agent eventually converges to a policy yielding:

$$\mathbb{E}_{x_1, x_2}[Y] = \mathbb{E}[x_2 \oplus U_2] = 0.5x_2 + 0.5(1 \oplus x_2) = 0.5.$$

Despite its naive formulation, the *all at once* strategy where the agent intervenes on all variables simultaneously leads to a policy that performs no better than random, achieving an expected reward of at most 0.5 regardless of the specific values of X_1 and X_2. Initially, one might attribute this limitation to sample complexity, suggesting that increased interactions would eventually lead to convergence. However, this strategy fundamentally disregards the causal structure $\mathcal{G}$, and no amount of sampling can compensate for this.

In contrast, consider a policy from the mixed policy space that intervenes only on X_1. The expected reward becomes:

$$\mathbb{E}_{x_1}[Y] = x_1,$$

implying that the optimal intervention is $\text{do}(X_1 = 1)$, which deterministically yields $Y = 1$.

This contrast reveals that the all at once strategy fails to converge, as it omits partial interventions entirely. Meanwhile, a *brute force* strategy that exhaustively explores all subsets of $\{X_1, X_2\}$ can recover the optimal policy. Empirical results (see Fig. 29d) show that the all at once agent incurs linear regret, while the brute force agent successfully converges.

Further insight is gained via Rule 3 of do calculus. Notably, $\mathbb{P}(y \mid \text{do}(x_1, x_2)) = \mathbb{P}(y \mid \text{do}(x_2))$ because X_1 does not affect Y under intervention on X_2. Thus, we have:

$$\mu_{x_1,x_2} = \mu_{x_2}.$$

To summarize, the expected rewards for selected interventions in the mixed policy space are:

$$\mathbb{E}[Y] = \mathbb{E}[X_1] = 0.5, \tag{17}$$
$$\mathbb{E}_{x_1}[Y] = x_1, \tag{18}$$
$$\mathbb{E}_{x_2}[Y] = 0.5x_2 + 0.5(1x_2) = 0.5. \tag{19}$$

These results highlight that partial interventions, guided by causal structure, can outperform strategies that ignore underlying causal mechanisms. Therefore, the optimal action for the original model $\mathcal{M} \rightarrow$ is $\text{do}(X_1 = 1)$, yielding an expected reward of $\mu_{X_1 \to 1} = 1$. Any joint intervention on $\{X_1, X_2\}$ results in suboptimal performance, incurring regret and preventing convergence to the optimal solution.

To analyze this further, we explore how interventional probabilities relate within the causal model from Fig. 29a. The observational distribution $\mathbb{P}(y) = \mathbb{P}(y \mid \text{do}(\emptyset))$ can be expressed as a convex combination of interventional

distributions:

$$\mathbb{P}(y) = \sum_{x_1} \mathbb{P}(y \mid x_1)\mathbb{P}(x_1) = \sum_{x_1} \mathbb{P}_{x_1}(y)\, \mathbb{P}(x_1). \tag{20}$$

Hence, the expected reward is:

$$\mu^{\emptyset} = \sum_{x_1} \mu_{x_1} \mathbb{P}(x_1) \leq \sum_{x_1} \mu^{\rightarrow}_{X_1} \mathbb{P}(x_1) = \mu^{\rightarrow}_{X_1}. \tag{21}$$

This inequality holds for any SCM compatible with Fig. 29a and emphasizes that interventions on non optimal arms (e.g., $\text{do}(X_2)$ or $\text{do}(X_1, X_2)$) lead to regret. Between these, intervening on X_2 is preferable over $\{X_1, X_2\}$ due to fewer required arms. Ultimately, $\text{do}(X_1)$ is preferable over doing nothing, as the observational reward $\mu^{\emptyset}$ is bounded above by $\mu^{\rightarrow}_{X_1}$.

A natural question is whether $\mu^{\rightarrow}_{X_2}$ can exceed $\mu^{\rightarrow}_{X_1}$ in some causal model. Indeed, this is possible. Consider a variant SCM $\mathcal{M}^{\star}$ with the same structure as Eq. (306), except the mechanism for Y is modified as:

$$f_Y := X_2 + U_{X_2,Y}. \tag{22}$$

Then, the expected rewards become:

$$\mathbb{E}[Y] = \mathbb{E}[(X_1 \oplus U_{X_2,Y}) + U_{X_2,Y}] = 1, \tag{23}$$

$$\mathbb{E}_{x_1}[Y] = 0.5x_1 + 0.5(2x_1) = 1, \tag{24}$$

$$\mathbb{E}_{x_2}[Y] = x_2 + 0.5. \tag{25}$$

This demonstrates that in $\mathcal{M}^{\text{`}}$, $\mu^{\rightarrow}_{X_2} > \mu^{\rightarrow}_{X_1}$, showing that the relative optimality of intervention sets depends on the SCM's specific mechanisms. Hence, the optimal action is $\text{do}(X_2 = 1)$, achieving an expected reward of $\mu_{X_2 \to 1} = 1.5$. This highlights the inherent limitation of relying solely on the causal diagram to determine the preferable interventional scope. The optimal strategy ultimately depends on the specific instantiation of the structural causal model (SCM), rather than just its graphical representation. In light of the example, it becomes clear that disregarding the underlying causal structure—specifically the relationship between the action space and reward—can lead to suboptimal decisions due to repeatedly selecting regret incurring arms. For instance, naive strategies that either intervene on all variables, $\text{do}(x_1, x_2)$, or on those closest to the outcome, such as $\text{do}(x_2)$, may never converge to the truly optimal arm, e.g., $\text{do}(x_1^{\rightarrow})$.

This example reveals the presence of equivalence classes of actions with respect to their expected rewards, as well as partial orderings among subsets of action variables. Notably, the expected reward of one action can often

be expressed using probabilities associated with other interventions. For example, the observational distribution $P(y)$ can be decomposed as

$$P(y) = \sum_{x_1} P_{x_1}(y)P(x_1) = \sum_{x_1} P_{x_1}(y)P_{x_2}(x_1).$$

Figure 30 illustrates the relationships among various interventional distributions and their rewards. It shows that certain distributions, like $P_{x_1}(y, x_2)$, can be derived from others, such as $P(y, x_1, x_2)$. Furthermore, the expected observational reward can be rewritten using expressions that depend on other arms' distributions. As we will see, such decompositions can significantly enhance the performance of online learning algorithms.

These ideas are explored more formally in the work of Lee and Bareinboim (2019a, 2018a, 2020).

5.1.1 Structural Properties in Mixed Policy Learning in a MAB Setting

In this section, we outline three structural properties that arise in bandit problems under mixed policy settings. These properties emerge due to shared causal mechanisms across actions and can be analyzed using tools like do calculus.

Property 1: Equivalence Among Actions. Do calculus provides a principled way to determine equivalences among interventional distributions. These equivalences effectively partition the action space into equivalence classes. Rule 3 of do calculus is particularly useful for identifying when an intervention does not affect the outcome variable. Specifically, if the following condition holds:

$$P(y \mid \text{do}(x, z)) = P(y \mid \text{do}(x)),$$

then we can conclude that the variable z has no causal effect on y when x is intervened upon. This can be verified graphically by checking if $Y \perp Z \mid X$ in the manipulated graph $G_{\overline{X}\underline{Z}}$.

In such cases, $\mu_{x,z} = \mu_x$, meaning interventions that differ only in z yield the same expected reward. Hence, in an online learning context where the goal is to minimize cumulative regret, it suffices to explore only a representative action from each equivalence class—particularly those containing the optimal arm.

Motivated by this observation, we introduce the concept of a *minimal intervention set*, which includes only the necessary interventions to identify the best arm without redundant exploration.

Definition 13 (Minimal Intervention Set (MIS))
A subset of action variables $X^\star \subseteq X^\vartheta$ is said to be a minimal intervention set *relative to a causal graph G, intervenable variables X^ϑ, and a target variable Y, if there exists no proper subset $X^{\text{“}} \subset X^{\text{‘}}$ such that*

$$\mu_{x^{\text{“}}} = \mu_{x^{\text{‘}}} \quad \textit{for every SCM conforming to } G,$$

and for every $x^{\text{“}} \in \mathcal{D}(X^{\text{“}})$ consistent with $x^{\text{‘}}$.

Proposition 3 (Minimality)
A set of variables $X^{\text{‘}} \subseteq X^\star$ is a minimal intervention set for G with respect to Y if and only if $X^\star \subseteq an(Y)_{G_{X^\star}}$.

This characterization demonstrates that one can consider only directed edges among variables (excluding unobserved variables) to acquire minimal intervention sets (MISes). Intervening on nothing or a single ancestor of Y in G constitutes MISes. Moreover, intervening on $W = \text{pa}(Z) \setminus Z$ is also an MIS for any $Z \subseteq \text{an}(Y)_G$, since each variable $W_i \in W$ has a directed path towards Y without passing through the rest of W (i.e., $W \setminus \{W_i\}$).

Property 2. Partial orders among minimal intervention sets.
Given a causal diagram G, there may exist partial orders among MISes where some sets of interventions are always at least as good as others across all SCMs consistent with G. Specifically, for two sets of variables $W, Z \subseteq X^\star$, it may hold that:

$$\max_{w \in \mathcal{D}(W)} \mu_w \leq \max_{z \in \mathcal{D}(Z)} \mu_z$$

for every SCM conforming to G. If this inequality holds, then actions over $\mathcal{D}(W)$ are redundant for identifying optimal interventions.

Definition 14 (Possibly Optimal Minimal Intervention Set (POMIS))
Given a causal diagram G, a set of intervenable variables $X^\star$, and target variable Y, let $X^{\text{‘}} \subseteq X^\star$ be a minimal intervention set (MIS). If there exists an SCM conforming to G such that

$$\mu_{x^{\text{‘}*}} > \max_{Z \in \mathcal{Z} \setminus \{X^\star\}} \mu_{z^*}$$

where $\mathcal{Z}$ is the collection of MISes and $x^{\text{‘}*}, z^*$ denote the optimal interventions over $X^\star, Z$ respectively, then $X^\star$ is a possibly optimal minimal intervention set (POMIS) with respect to G, $X^\star$, and Y.

Definition 15 (Possibly Optimal Minimal Intervention Set (POMIS))
Given a causal diagram G, a set of intervenable variables $X^\star$, and a target

variable Y, let $X^\star \subseteq X^\vartheta$ be a minimal intervention set (MIS). If there exists an SCM conforming to G such that

$$\mu_{x^{`*}} > \max_{Z \in \mathcal{Z} \setminus \{X^\star\}} \mu_{z^*},$$

where $\mathcal{Z}$ is the set of all MISes with respect to G, X^ϑ, and Y, and $x^{`*}, z^*$ denote the optimal interventions over $X^\star$ and Z, respectively, then $X^\star$ is a possibly optimal minimal intervention set (POMIS).

To determine whether a subset of X^ϑ is a POMIS, one could enumerate all possible partial orders among MISes and select those not dominated by any other. However, it remains unclear under what graphical conditions two arbitrary MISes can be compared or whether such conditions are complete.

As a starting point, we consider a partial order between two MISes using Rule 2 of do calculus. Let W be an MIS and consider an intervention on W. Then, for some $Z \subseteq X^\vartheta$:

$$\mathbb{E}_w[Y] = \sum_z \mathbb{E}_w[Y \mid z]\, P_w(z).$$

If Rule 2 allows converting the observation z into an intervention, this becomes:

$$\mathbb{E}_w[Y] = \sum_z \mathbb{E}_{z,w}[Y]\, P_w(z) \leq \sum_z \mathbb{E}_{(z,w)^*}[Y]\, P_w(z) = \mathbb{E}_{(z,w)^*}[Y] = \mathbb{E}^*_{z^{`},w^{`}}[Y],$$

where $Z^{`} \cup W^{`}$ is the MIS corresponding to $Z \cup W$ and $(z,w)^*$ denotes the assignment that maximizes the expectation.

To identify such $Z \subseteq X^\vartheta$, we use Rule 2, which requires:

$$(Y \perp Z \mid W)_{G_{W,Z}} \quad \text{or equivalently,} \quad (Y \perp Z)_{(G \setminus W)_Z}.$$

If no backdoor path from Z to Y exists in $G \setminus W$, then:

$$\mu^*_W \leq \mu^*_{W,Z} = \mu^*_{W^{`},Z^{`}}.$$

This process can be repeated iteratively until no non intervened subset of X^ϑ contains a backdoor path to Y.

Figure 31 illustrates this logic. In the base graph (Fig. 31a), each X_i for $i > 1$ has a backdoor path via Z, and X_4 is also directly confounded with Y. After intervening on X_1 (Fig. 31b), the light blue area shows all remaining backdoor paths to Y in G_{X_1}, implying:

$$\mu^\emptyset \leq \mu_{x_1}.$$

In simplified projections, we ignore variables not in X or Y, yet the same backdoor paths remain visible. Upon intervening on X_2, X_1 and X_3 lose their backdoor connections to Y. Then, due to minimality, we conclude:

$$\mu^*_{X_2} \leq \mu^*_{X_2,X_3,X_1} \leq \mu^*_{X_3} \quad .$$

We may exclude variables ineffective to Y under $\text{do}(x_2)$ initially. To that end, we define a set of variables with backdoor paths to Y in a given graph:

5.2 Structural Properties in Mixed Policy Learning in a MAB Setting

The subgraph induced by the Minimal Unobserved Confounders' Territory (MUCT) represents how Y is determined through the variables ruled by unobserved confounders under interventions outside the MUCT. MUCT is tightly related to Rule 2 of do calculus, where the procedure iteratively expands variables that have backdoor paths to Y, such that the remaining variables become exchangeable between conditioning and intervention.

This relates to the partial orders we aim to establish. Specifically, this implies that:

$$\mu_{\emptyset} \leq \mu_{\to V \setminus T} = \mu_{\text{pa}(T) \setminus T}$$

where $X_\vartheta = V \setminus \{Y\}$. We define:

- $\text{MUCT}(G, Y)$ as the Minimal UC Territory of G with respect to Y,
- $\text{IB}(G, Y) = \text{pa}(T)_G \setminus T$ as the *Interventional Border* of G with respect to Y, where $T = \text{MUCT}(G, Y)$.

Given an arbitrary $X_\vartheta \subseteq V \setminus \{Y\}$, one can obtain a MUCT and IB from H, the latent projection of G onto $X \cup \{Y\}$. Then:

$$\mu_{\to W} \leq \mu_{\to \text{IB}(H_W, Y)}$$

A more involved example is depicted in Figure 32, where the MUCT can be constructed by iteratively updating $\{Y\}$ to include variables confounded with Y, their descendants, and confounded variables of those descendants, e.g., yielding $\{Y, C, E, F\}$ under $do(B)$.

Theorem 7 (Lee and Bareinboim, 2018a, 2019a)
Given G, X_ϑ, and Y, let $X^{`} \subseteq X_\vartheta$ be an MIS and let $H$ be the latent projection of $G$ onto $X \cup \{Y\}$. Then, $X^{`}$ is a POMIS if and only if

$$\text{IB}(H_{X^{`}}, Y) = X^{`}$$

This result can be interpreted as follows: under the distributions induced by the unobserved confounders in a MUCT, the internal mechanisms within the MUCT are finely tuned to yield optimal outcomes for some external configurations. Intervening on any of the variables in the MUCT disrupts this mechanism, often yielding suboptimal rewards. For example, in Figure 31, the arms $\{X_1\}$, $\{X_3\}$, and $\{X_4\}$ are identified as POMISes.

Property 3. Expressions Among Actions The first two structural properties guide bandit agents to focus on a reduced set of possibly optimal arms before observing any data. The third property connects the causal effect of playing an arm with the distributions obtained through other arms.

Consider the causal diagram in Figure 33 with $X_\vartheta = \{B, D, E\}$ where the POMISes include $\emptyset$, $\{B\}$, $\{D\}$, $\{E\}$, $\{B, D\}$, and $\{D, E\}$. An agent playing only POMIS arms will observe samples from distributions such as $P(V)$, $P_b(V \setminus \{B\})$, ..., $P_{d,e}(V \setminus \{D, E\})$.

By c factorization, the expected reward under $do(\emptyset)$ is:

$$\mathbb{E}[Y] = \sum_{a,b,c,d,e,y} y \cdot P_{d,e}(y, a, b) \cdot P_{a,c}(e) \cdot P_c(d) \cdot P_b(c) \tag{26}$$

$$= \sum_{a,b,c,d,e,y} y \cdot P_{d,e}(y \mid a, b) \cdot P_{d,e}(a, b) \cdot P_{a,c}(e) \cdot P_c(d) \cdot P_b(c) \tag{27}$$

Each term above can be rewritten or estimated using data from other arms using do calculus (see Table 15). For instance, $P_c(d)$ can be estimated from a weighted combination of terms from different arms:

$$\hat{P}_c(d) = \frac{N_c \hat{P}_c(d) + N_{\emptyset|c} \hat{P}(d \mid c) + \sum_e N_{e|c} \hat{P}_e(d \mid c) + \sum_b N_{b|c} \hat{P}_b(d \mid c)}{N_c + N_{\emptyset|c} + \sum_e N_{e|c} + \sum_b N_{b|c}}$$

where $N_{w|z}$ is the number of samples with $Z = z$ in the distribution $do(W = w)$.

This estimator is based on maximum likelihood, aggregating compatible observations from different arms. By plugging these estimators into the expression for $\hat{\mathbb{E}}[Y]$, one obtains a more data efficient estimator.

In practice, this means that an online agent can confidently estimate $\mu_\emptyset = \mathbb{E}[Y]$ without excessively sampling the suboptimal arm associated with $do(\emptyset)$, avoiding high regret. For instance, as proposed in , bootstrapping is used to estimate the posterior of the expected reward for each POMIS, which is then integrated into Thompson Sampling. Similarly, variance estimates from bootstrap samples can be translated into effective counts for use in UCB based algorithms.

5.3 Mixed Policy with context

5.3.1 Causal Decision Making Over Mixed Policies with Contexts

As discussed in the previous section, a causal understanding of the environment enables agents to reason over a broad range of policies across various policy spaces, allowing them to decide not only which variables to intervene on, but also which to observe as part of their context. In this section, we explore systematic decision making over mixed policies where contexts are involved.

To illustrate the concept, consider an agent operating in the environment.

5.4 Contextual Minimality in Optimal Mixed Policies

Optimizing a mixed policy involves evaluating whether each component of the policy space (actions or contexts) contributes effectively. An action on X is considered meaningful if it affects Y through a change in its mechanism π_X. A context $S \in S_X$ is considered relevant if it provides information relative to other contexts $S_X \setminus \{S\}$.

However, this basic characterization is often insufficient when evaluating subsets of contexts across multiple actions, especially under optimized mechanisms π^{Π}.

The key idea is to identify contexts that can be inferred (or fixed) using the rest of the context set. These relationships are captured more formally in Equation 320. The fixed values do not need to be contexts themselves; any variable whose value can be inferred from others can be fixed to simplify the policy expression.

For our purposes, we aim to rewrite expected reward expressions such that any fixed context is inferable from remaining context variables.

Example: Fixing for Simplification Consider:

$$\sum_{a,b,d} P(a \mid b, c) f(a, b, d) \leq \sum_{b,d} f(a \to (b), b, d) = \sum_{b,d} f^{`}(b, d)$$

With C fixed, we say a is fixed conditional on B. The corresponding decision rule f drops a by inferring it from b and c.

Definition 16 (Minimality under Optimality)
*Given a causal graph G, target variable Y, decision variables X_ϑ, and context variables S_ϑ, a policy space Π is **minimal under optimality** if there exists a structural causal model $\mathcal{M}$ compatible with G such that:*

$$\mu_{\to}^{\Pi} > \mu_{\to}^{\Pi_{`}} \quad \textit{for all } \Pi_{`} \subset \Pi$$

That is, any strictly smaller policy space leads to strictly worse performance in some scenario:

$$\mathcal{R}_{\mathcal{M}\models G} \to \Pi_{\cdot} \subset \Pi \Rightarrow \mu_{\to}^{\Pi} > \mu_{\to}^{\Pi_{\cdot}}$$

Next, we will derive a sufficient condition for detecting non minimality under optimality by:

1. Constructing an intermediate expected reward expression given a candidate set of variables to fix.

2. Checking whether this expression leads to a valid, simpler policy space.

Step 1: Deriving an Intermediate Expression from the Expected Reward

Consider a policy space Π that meets the basic minimality criteria. Let $X' \subseteq X(\Pi)$ be a subset of actions of interest (i.e., actions we may intervene upon), and $S' \subseteq S_X \setminus X'$ represent non action variables within the context (i.e., those we hold fixed).

Given:

(i) a subset $U' \in \mathcal{G}_{\Pi}^{\mathbb{A}}$ of exogenous variables,

(ii) a subset Z of endogenous variables disjoint from $S' \cup X'$ such that $Z \subseteq S_X \setminus (S' \cup X')$, and

(iii) an ordering $\prec$ over $V' \triangleq S' \cup X' \cup Z$,

such that the triple $(U', Z, \prec)$ satisfies certain structural constraints (see Lee and Bareinboim, 2020, Lemma 1), the intermediate reward expression μ_π can be written as:

$$\mu_\pi = \sum_{u'} P(u') \sum_{y,s',x'} y\, Q_\pi(y, s' \mid x') \sum_{z} \prod_{Z \in \mathcal{Z}} P_\pi(z \mid v'_z, z_{\prec Z}, u') \prod_{x \in X'} \pi(x \mid s_X \backslash \{x\}, z),$$

where $Q_\pi = P_{\pi, X'}$ is the distribution governed by π except over X'. The conditions are engineered to isolate the term $Q_\pi(y, s' \mid x')$ such that Z becomes irrelevant and does not influence it, allowing us to simplify the expected reward by moving to a more minimal policy that only considers context S'.

To elaborate, we substitute $P_\pi(z \mid V'_{\prec Z}, u')$ with $P_\pi(z \mid V_Z)$, where V_Z is the smallest subset of $V'_{\prec Z} \cup U'$ that suffices for determining Z. The goal is to reparameterize the expectation into $\mu_{\pi'}$, where decisions $(\pi'(X \mid S_X))$ are

made solely over X', using the transformation $\pi(x \mid S_X \setminus \{x\}, z) \to \pi(x \mid S_X \setminus \{x\})$. This is achieved by appropriately fixing the values of variables in the contexts needed for Z (those in V_Z) so that the conditional terms $P_\pi(z \mid v_Z)$ become analogous to $P(z \mid c_1)$ as encountered earlier.

We observe that identifying and fixing unnecessary contexts simplifies this process and eliminates those not essential. This provides insight into the careful selection of variables—choosing contexts and exogenous variables wisely avoids overcomplication. The ordering $\prec$ ensures a valid factorization of the joint probability distribution and indicates the relative ordering of variable fixation.

Step 2: Converting the Intermediate Expression to the Expected Reward under a Simpler Policy

After obtaining the intermediate expression for some $(U', Z, \prec)$, we convert this into the simplified expression $\mu_{\pi'}$, where the revised policy space is defined as $\Pi' \triangleq (\Pi \setminus X') \cup \{(x, S_X \setminus Z)\}_{x \in X'}$. This conversion is accomplished by fixing Z values unconditionally.

During this transformation:

- Each variable in Z whose ancestral variables lie entirely in $V'_{\prec Z}$ is fixed as it only depends on contexts, decisions, and exogenous variables.
- Fixing each z at the value which maximizes the ability to eliminate the term $P(z \mid \cdot)$ simplifies the overall expression.

This leads to a streamlined policy representation relying solely on the variables X' and context S'.

Constructing a Simpler Policy via a Dependency Graph

The process of deriving a simpler policy from the intermediate reward expression can be framed as constructing and analyzing a *dependency graph*. This graph, structured as a Directed Acyclic Graph (DAG), is based on the conditional distributions over Z and $U_\cdot$.

Graph Construction. We begin by introducing nodes for all variables in $U_\cdot$, Z, the parents of Z, and the parents of the intervened variables. Directed edges are then drawn according to the structure of the conditional distributions. For instance, if a distribution $P_\pi(Z \mid V_Z)$ exists, then the node Z will have directed edges from all variables in V_Z, establishing them as its parents. This rule also applies to decision functions, such as $\pi(x \mid s_X)$, where the parents of x in the graph are given by the contexts in s_X.

Marking Nodes to Simplify Policy. Once the DAG is constructed, we assess whether all variables in Z can be fixed to derive a simpler policy. To begin, identify the subset of context variables $S_{\cdot}$ that have no parents in the graph; these can be marked immediately as they are available to condition on other variables. For example, in the provided case, C_1 can be used to determine values for C_2.

Nodes are then marked iteratively: a node can be marked once all its parent nodes are marked. For instance, C_3 may be marked unconditionally if it has no parents. Then, X_2 can be marked provided C_3 is already marked. Together with C_1 and C_3, we can fix C_2 using:

$$c_2(c_3, x_2(c_3), c_1) = c_2(c_1),$$

implying the other dependencies can be simplified away.

Conditions for Policy Simplification. We can finalize the simplified policy if the following two conditions are met:

1. All variables in Z are successfully marked (i.e., fixed).
2. For each intervened variable X, its ancestors do not include any context variables that will be unavailable in the simplified policy, i.e., variables in $S_{\cdot} \setminus (S_X \setminus Z)$.

When both criteria are satisfied, the intermediate expression can be reduced to an expected reward expression under a simpler policy.

6 Counterfactual Decision Making

In this section, we explore a novel mode of interaction between the agent and the environment through **Layer 3** of the *Probability Causality Hierarchy* (PCH). Previous tasks relied on interactions characterized by passive observation (*seeing*) and active experimentation (*doing*), corresponding to Layers 1 and 2 of the PCH. Here, we introduce an advanced interaction modality, enabling agents to reason over a space of **counterfactual policies**, denoted ΠCTF.

Task Signature: Counterfactual Randomization

We define an online learning task with *counterfactual randomization* using the following signature:

$$T_{\text{ctf rand}} = \langle I = \text{ctf},\ A = \mathcal{A},\ \Pi = \Pi\text{CTF},\ R = \mathcal{D}(Y) \to \mathbb{R} \rangle.$$

This task seeks a policy π^* such that:

$$\pi^* = \arg \max_{\pi \in \Pi\text{CTF}} \mathbb{E}_{\pi}^{M} \left[R(Y) \right],$$

where the defining feature is the use of *counterfactual interactions* to guide learning.

Counterfactual vs. Interventional Interaction

To contextualize this setting, recall the interventional policy space Π_{EXP} . When an agent follows a policy $\pi \in \Pi_{\text{EXP}}$ and performs interventions $\text{do}(\pi)$, it induces an **interventional distribution** $P_{\pi}(V)$, typically grounded in Fisherian randomization.

In contrast, when the agent follows a policy $\pi \in \Pi_{\text{CTF}}$, it engages with the environment through **counterfactual interactions**, gaining access to x *specific counterfactual effects* of decisions on outcomes. These counterfactual effects characterize a unique distribution type, described in detail in Plecko and Bareinboim (2022, Sec. 4.1.1), and enable decision making informed by "what would have happened if..." reasoning.

Counterfactual Randomization (ctf rand)

We will introduce a specific form of counterfactual interaction termed *counterfactual randomization* (ctf rand), which enables agents to act optimally

even in complex decision making environments where interventional approaches fall short. This new methodology expands the agent's learning capabilities by leveraging the structure of counterfactuals for policy optimization.

6.1 Counterfactual Decision Criterion

We begin by outlining how an agent interacts with a multi armed bandit (MAB) environment under different causal regimes.

In the **observational** setting, the agent passively observes actions and outcomes, both influenced by unobserved confounders U. The system behavior is captured by a policy f_X, which maps U to the arm choice X, creating a spurious association between X and the reward Y. This is represented by a bidirected arrow $X \dashleftarrow\dashrightarrow Y$, which reflects natural biases or *predilections* for example, gamblers preferring blinking machines or doctors relying on intuition. The expected reward in this regime is $\mathbb{E}[Y \mid x]$.

In the **interventional** setting, these natural biases are overridden. An external policy (e.g., random assignment via coin flips) enforces actions through $\text{do}(x)$, breaking the influence of U on X. The resulting causal graph removes the bidirected edge between X and Y. The expected reward is now $\mathbb{E}_x[Y]$, determined by the reward function $f_Y(X, U)$ and the exogenous distribution $P(U)$.

The observational agent operates within its inherent preferences, while the interventional agent explores the environment independently. By leveraging both regimes and their respective distributions, we can understand and improve the agent's natural decision making tendencies through counterfactual reasoning.

We begin by outlining how an agent interacts with a multi armed bandit (MAB) environment under different causal regimes.

In the **observational** setting, the agent passively observes actions and outcomes, both influenced by unobserved confounders U. The system behavior is captured by a policy f_X, which maps U to the arm choice X, creating a spurious association between X and the reward Y. This is represented by a bidirected arrow $X \dashleftarrow\dashrightarrow Y$, which reflects natural biases or *predilections* for example, gamblers preferring blinking machines or doctors relying on intuition. The expected reward in this regime is $\mathbb{E}[Y \mid x]$.

In the **interventional** setting, these natural biases are overridden. An external policy (e.g., random assignment via coin flips) enforces actions through $\text{do}(x)$, breaking the influence of U on X. The resulting causal graph removes the bidirected edge between X and Y. The expected reward is now $\mathbb{E}_x[Y]$, determined by the reward function $f_Y(X, U)$ and the exogenous distribution $P(U)$.

The observational agent operates within its inherent preferences, while the interventional agent explores the environment independently. By leveraging both regimes and their respective distributions, we can understand and improve the agent's natural decision making tendencies through counterfactual reasoning.

Definition 17 (Counterfactual Distribution (Bareinboim et al., 2020))
A Structural Causal Model (SCM) $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P \rangle$ defines a family of joint distributions over counterfactual events such as $Y_x, \ldots, Z_w$, for any variables $Y, Z, \ldots, X, W \in \mathbf{V}$.

The counterfactual distribution is given by:

$$P(y_x, \ldots, z_w) = \sum_u \mathbb{1}\{Y_x(u) = y, \ldots, Z_w(u) = z\} \cdot P(u),$$

where the indicator function selects worlds where the counterfactual variables take the specified values under interventions $x, w, \ldots$ and $P(u)$ is the distribution over exogenous variables.

Example 2 (Evaluating Counterfactual Distributions)
Note that in Equation above, the left hand side includes variables with differing subscripts, each representing a distinct counterfactual world. The evaluation of this expression can be understood through the following procedure:

1. *For each group of subscripts (e.g., $do(x), \ldots, do(w)$ applied to $Y, \ldots, Z$ respectively), replace the corresponding structural mechanisms with constants to obtain modified functions $F_x, \ldots, F_w$, thus defining submodels $\mathcal{M}_x, \ldots, \mathcal{M}_w$ of the original SCM $\mathcal{M}$.*

2. *For every possible setting of exogenous variables $U = u$, evaluate the altered mechanisms in a valid topological order—ensuring that any variable on the left hand side is computed only after all variables on the right hand side of its structural equation have been evaluated. This yields the counterfactual outcomes of the observed variables.*

3. *Finally, accumulate the probability mass $P(U = u)$ for all instantiations of u that are consistent with the specified counterfactual events (e.g., $Y_x = y, \ldots, Z_w = z$), which correspond to $Y = y, \ldots, Z = z$ in the submodels $\mathcal{M}_x, \ldots, \mathcal{M}_w$.*

The counterfactual analysis outlined earlier motivates a novel decision making principle in MAB environments. Rather than comparing expected rewards under interventions (i.e., $\mathbb{E}_x[Y]$ for each action x), the agent should

compare the expected outcomes of acting in line with or against their natural inclination. Formally, the counterfactual policy is given by:

$$\pi^*(x) = \arg\max_x \mathbb{E}\left[Y_{X \leftarrow x} \,\middle|\, X = x^{\text{`}}\right], \tag{28}$$

where $x^{\text{`}}$ denotes the agent's natural (observed) choice, and x is a candidate action. This policy, called the *Counterfactual Decision Criterion* (CDC), compares potential rewards conditioned on the agent's original inclination, thus incorporating personalized information encoded in the confounders.

For instance, in the binary case where the observed choice is $X = 1$, the agent evaluates:

$$\mathbb{E}[Y_{X \leftarrow 0} \mid X = 1] \quad \text{vs.} \quad \mathbb{E}[Y_{X \leftarrow 1} \mid X = 1].$$

If the counterfactual reward of choosing $X = 0$ is higher, the agent should act *against* their intuition and choose $X = 0$; otherwise, they should follow it. This introduces a new class of *counterfactual policies* that optimize not just over rewards but over personalized, intuitive deviations.

Definition 18 (Counterfactual Policy (MAB))
Let M be a multi armed bandit (MAB) environment. A counterfactual policy space Π_{CTF} is defined as a set of policies $\pi(X \mid X^\star)$ that map an intended action $X^\star$ to a probability distribution over possible realized actions X. We will denote this policy space by:

$$\Pi_{\text{CTF}} = \{\pi(X \mid X^\star)\}.$$

At first glance, the structure of a counterfactual policy $\pi(X \mid X^\star)$ might seem unintuitive, as both its input and output are defined over the domain of the action variable X. This appearance of self reference could suggest a circular dependency in the system. However, from a semantic standpoint, $\pi(X \mid X^\star)$ is a well defined and computable counterfactual quantity in any structural causal model (SCM) M.

Although $X^\star$ and X belong to the same action space, they represent variables from distinct hypothetical settings. The intended action $X^\star$ corresponds to the agent's inherent behavioral tendency, as seen in the observational distribution from layer 1 , whereas the realized action X represents the outcome of an intervention, generating the interventional distribution observed in layer 2.

Choosing an intervention $do(X \leftarrow x)$ in accordance with the agent's predisposition $X = x^{\text{`}}$ results in the expected reward associated with the counterfactual policy $\pi(X \mid X^\star)$.

Definition 19 (Counterfactual Submodel MAB)
Let M be a multi armed bandit (MAB) environment as depicted in Fig. 43a, described by the structural causal model:

$$M = \langle U = \{U\},\ V = \{X, Y\},\ F = \{f_X, f_Y\},\ P = P(U)\rangle.$$

Let $\pi(X \mid X^\star)$ be a counterfactual policy defined over the action variable X. A corresponding counterfactual submodel M_π of M is defined as a modified SCM:

$$M_\pi = \langle U = \{U\},\ V_\pi = \{X^\star, X, Y\},\ F_\pi,\ P = P(U)\rangle,$$

where the structural functions F_π are specified by:

$$F_\pi = \begin{cases} X^\star \coloneqq f_X(U), \\ X \sim \pi(X \mid X^\star), \\ Y \coloneqq f_Y(X^\star, U). \end{cases}$$

Figure 43c presents a graphical depiction of the submodel M_π, which is induced by a counterfactual policy $\pi(X \mid X^\star)$. In this construction, a new virtual variable $X^\star$ is introduced to represent the agent's intended action. This variable serves as an input to determine the realized action X, which in turn influences the reward variable Y.

Formally, the expected reward $E_\pi[Y; M]$ under a counterfactual policy $\pi(X \mid X^\star)$ in an MAB environment is defined as the expected value of Y in the counterfactual submodel M_π:

$$E_\pi[Y; M] = \sum_{x^{\text{`}}, x} \mathbb{E}\left[Y \mid x, x^{\text{`}}; M_\pi\right] \cdot P(x \mid x^{\text{`}}; M_\pi) \cdot P(x^{\text{`}}; M_\pi) \tag{29}$$

$$= \sum_{x^{\text{`}}, x} \mathbb{E}\left[Y \mid x, x^{\text{`}}; M_\pi\right] \cdot \pi(x \mid x^{\text{`}}) \cdot P(x^{\text{`}}). \tag{30}$$

The second equality follows from the definition of the submodel M_π (as per Eq. 348), where $X^\star$ is generated via the behavioral policy f_X from the original model M, and the realized action X is sampled from the counterfactual policy π. Since π does not depend on the unobserved confounder U, conditioning on $X^\star$ effectively blocks any backdoor paths from X to Y. Consequently, the joint conditioning on both the intended and realized actions $(X^\star, X)$ in M_π recovers the counterfactual effect of treatment on the treated (ETT). The expression can thus be rewritten as:

$$E_\pi[Y; M] = \sum_{x, x^{\text{`}}} \mathbb{E}\left[Y_{X:=x} \mid X = x^{\text{`}}\right] \cdot \pi(x \mid x^{\text{`}}) \cdot P(x^{\text{`}}).$$

More generally, observe that the counterfactual policy space Π_{CTF} subsumes the experimental policy space $\Pi_{\text{EXP}} = \{\pi(X)\}$ in MAB environments. For any experimental policy $\pi(X)$, one can emulate its behavior using a counterfactual policy $\pi(X \mid X^\star)$ by setting the realized action via the intervention $\text{do}(X \coloneqq x)$ independently of the agent's natural inclination $X^\star = x^{\text{`}}$. That is,

$$\pi(x) = \pi(x \mid x^{\text{`}}), \quad \forall x, x^{\text{`}}.$$

Under this formulation, the expected reward simplifies to:

$$E_\pi[Y] = \sum_x \pi(x) \sum_{x^{\text{`}}} \mathbb{E}\left[Y_{X:=x} \mid X = x^{\text{`}}\right] \cdot P(x^{\text{`}}) \tag{31}$$

$$= \sum_x \pi(x) \cdot \mathbb{E}[Y_x]. \tag{32}$$

By the definition of potential outcomes and interventional distributions , the counterfactual expression $\mathbb{E}[Y_x]$ corresponds exactly to the interventional expectation $\mathbb{E}_x[Y]$. Hence, the above equation recovers the expected reward of an experimental policy $\pi(x)$.

As will be shown next, an optimal counterfactual policy can always match or outperform the best experimental policy in terms of expected reward.

Theorem 8 (Counterfactual Policies Dominate Interventional Policies MAB) *Let M be an MAB environment. Define the counterfactual policy space as $\Pi_{\text{CTF}} = \{\pi(X \mid X^\star)\}$ and the interventional policy space as $\Pi_{\text{EXP}} = \{\pi(X)\}$. Then, the optimal counterfactual policy achieves expected reward at least as high as the optimal interventional policy:*

$$\max_{\pi \in \Pi_{\text{CTF}}} \mathbb{E}_\pi[Y] \geq \max_{\pi \in \Pi_{\text{EXP}}} \mathbb{E}_\pi[Y].$$

A natural question arises: under what conditions does the equality in Theorem 15 hold, allowing a standard interventional agent to match the performance of an optimal counterfactual agent? This occurs when the *No Unobserved Confounding* (NUC) condition (Definition 13) holds in the underlying MAB environment. Under this condition, there is no unobserved confounder simultaneously influencing both the action X and the reward Y. As a result, the agent's intended action $X^\star$ is independent of the potential outcome Y_x induced by the intervention $\text{do}(X \leftarrow x)$:

$$X^\star \perp\!\!\!\perp Y_x.$$

When this independence holds, the expected reward under the counterfactual submodel M_π simplifies as follows:

$$\mathbb{E}[Y; M_\pi] = \sum_x \mathbb{E}[Y_{X:=x}] \sum_{x^{\text{`}}} \pi(x \mid x^{\text{`}}) \cdot P(x^{\text{`}}) \tag{33}$$

$$= \sum_x \mathbb{E}[Y_{X:=x}] \cdot P_\pi(x), \tag{34}$$

where $P_\pi(x) = \sum_{x^{\text{`}}} \pi(x \mid x^{\text{`}}) \cdot P(x^{\text{`}})$ denotes the marginal distribution over realized actions under the counterfactual policy π.

This result implies that an interventional agent can simulate the behavior of the counterfactual policy by adopting an experimental policy defined as $\pi(x) = P_\pi(x)$. Therefore, when the NUC condition is satisfied, the optimal counterfactual and experimental policies achieve identical performance.

6.1.1 Markov Decision Processes with Unobserved Confounders

In this section, we extend the notion of counterfactual policies to a broader class of sequential decision making problems, where an agent must determine a sequence of actions over time. Our analysis centers on a canonical class of environments—*Markov Decision Processes* (MDPs)—augmented through the framework of structural causal models (SCMs). This formulation generalizes classical MDPs (cf. Puterman, 1994) to accommodate settings with latent confounding, thereby enabling a principled treatment of unobserved factors influencing both actions and outcomes.

Definition 20 (MDP Environment)
An MDP environment *is represented as a structural causal model (SCM)*

$$M = \langle \mathbf{U}, \mathbf{V}, F, P(\mathbf{U}) \rangle,$$

where the decision horizon is $H \in \mathbb{N}_+$. The components are defined as follows:

- $\mathbf{U} = \{U_1, \ldots, U_H\}$ *is a sequence of exogenous variables;*
- $\mathbf{V} = \{S, X, Y\}$ *comprises endogenous variables consisting of sequences of states $S = \{S_1, \ldots, S_H\}$, actions $X = \{X_1, \ldots, X_H\}$, and rewards $Y = \{Y_1, \ldots, Y_H\}$;*
- *F is a collection of structural assignments, defined for each $i = 1, \ldots, H$ as:*

$$S_i \coloneqq f_S(S_{1:i1}, X_{1:i-1}, U_i),$$
$$X_i \coloneqq f_X(S_i, U_i),$$
$$Y_i \coloneqq f_Y(S_i, X_i, U_i);$$

- $P(\mathbf{U})$ *is a product distribution over the exogenous variables, given by* $P(\mathbf{U}) = \prod_{i=1}^{H} P(U_i)$*, where each* U_i *is drawn i.i.d. from a common distribution over domain* $\mathcal{D}(\mathbf{U})$*.*

Definition abo introduces a generalized class of environments analogous to Markov Decision Processes (MDPs), in which the No Unobserved Confounders (NUC) assumption does not hold. These environments, referred to as *MDPs with Unobserved Confounders (MDPUCs)*, have been formally studied in recent work by Zhang and Bareinboim in 2022.

In a standard MDP, the Markov property ensures that both the observational distribution $P(S, X, Y)$ and the interventional distribution $P_\pi(S, X, Y)$, induced by a policy $\pi = (\pi_1(X_1 \mid S_1), \ldots, \pi_H(X_H \mid S_H))$, satisfy the conditional independence:

$$\bar{S}_{i+1:H}, \bar{X}_{i:H}, \bar{Y}_{i:H} \perp\!\!\!\perp \bar{S}_{1:i}, \bar{X}_{1:i}, \bar{Y}_{1:i} \mid S_i, \quad \text{for all } i = 1, \ldots, H. \tag{35}$$

This condition can be verified graphically using d-separation in the causal diagram of the MDP.

However, in the MDPUC setting, the existence of unobserved confounders invalidates this equivalence. In particular, the transition dynamics and reward structure may differ between observational and interventional regimes. Specifically, for each time step i:

$$P(s_{i+1} \mid s_i, x_i) \neq P_{x_i}(s_{i+1} \mid s_i), \tag{36}$$

$$\mathbb{E}[Y_i \mid s_i, x_i] \neq \mathbb{E}_{x_i}[Y_i \mid s_i]. \tag{37}$$

These discrepancies underscore the breakdown of the standard MDP assumptions when latent confounding is present.

We now formalize the notion of counterfactual policies in the context of MDPUCs. Analogous to the multi-armed bandit (MAB) setting, an agent employing a counterfactual policy selects its action X_i based on its original intended action, as well as on counterfactual considerations.

Crucially, in the MDP setting, the agent also incorporates the observed current state S_i into its decision-making process. This distinguishes MDPUC counterfactual policies from those in MABs, where the state is static or absent. In this way, counterfactual reasoning allows agents to query alternative histories—i.e., "what would have happened if a different action were taken in this state?"—and exploit this richer informational structure to refine their decision policy over time.

The remainder of this section explores how counterfactual policies can be constructed and optimized in MDPUC environments, and introduces the causal machinery necessary to support this reasoning over trajectories.

Definition 21 (Counterfactual Policy – MDP)
For an MDP environment $\mathcal{M}$, a counterfactual policy space Π_{CTF} *is defined as a collection of policies:*

$$\pi = (\pi_1(X_1 \mid S_1, X_1^\star),\ \ldots,\ \pi_H(X_H \mid S_H, X_H^\star))\,,$$

where each decision rule $\pi_i(X_i \mid S_i, X_i^\star)$ maps from the joint domain of the observed state S_i and the intended action $X_i^\star$ to a probability distribution over the space of realized actions X_i.

We denote the counterfactual policy space compactly as:

$$\Pi_{CTF} = \{(X_i, \{S_i, X_i^\star\})\}_{i=1}^{H}\,.$$

Definition 22 (Counterfactual Submodel – MDP)
Let $\mathcal{M} = \langle U, V, F, P(U)\rangle$ be a Markov Decision Process (MDP) environment, and let $\pi = (\pi_1(X_1 \mid S_1, X_1^\star), \ldots, \pi_H(X_H \mid S_H, X_H^\star))$ be a counterfactual policy over the action sequence $X = \{X_1, \ldots, X_H\}$.

Then, the counterfactual submodel $\mathcal{M}_\pi$ *is defined as a Structural Causal Model (SCM):*

$$\mathcal{M}_\pi = \langle U, V_\pi = \{S, X^\star, X, Y\}, F_\pi, P(U)\rangle,$$

where: - $S = \{S_1, \ldots, S_H\}$ is the set of states, - $X^\star = \{X_1^\star, \ldots, X_H^\star\}$ is the set of intended actions, - $X = \{X_1, \ldots, X_H\}$ is the set of realized actions, - $Y = \{Y_1, \ldots, Y_H\}$ is the set of rewards, - F_π is a modified set of structural assignments:

$$F_\pi = \begin{cases} S_i \coloneqq f_S(S_{1:i-1}, X_{1:i-1}, U_i) \\ X_i^\star \coloneqq f_X(S_i, U_i) \\ X_i \sim \pi_i(X_i \mid S_i, X_i^\star) \\ Y_i \coloneqq f_Y(S_i, X_i, U_i) \end{cases} \qquad \text{for } i = 1, \ldots, H.$$

Counterfactual Markov Property. The causal diagram in Fig. 44b represents the submodel $\mathcal{M}_\pi$ induced by an MDP environment and a counterfactual policy $\pi(X_i \mid S_i, X_i^\star)$. The joint distribution $P_\pi(S, X^\star, X, Y)$ describes the probability distribution over the endogenous variables in $\mathcal{M}_\pi$.

By inspection, the data-generating mechanism defined by the causal structure in $\mathcal{M}_\pi$ gives rise to a Markov chain (Puterman, 1994). At each decision stage $i = 1, \ldots, H$, the state S_i and intended action $X_i^\star$ satisfy a conditional independence with respect to the past trajectory, which reflects a generalized Markov property in the counterfactual model:

$$P(S_{i+1}, X_{i+1}^\star \mid \bar{S}_{1:i}, \bar{X}_{1:i}^\star, \bar{X}_{1:i}; \mathcal{M}_\pi) = P(S_{i+1}, X_{i+1}^\star \mid S_i, X_i^\star, X_i; \mathcal{M}_\pi), \quad (38)$$

and similarly, the expected reward at each step i obeys the following conditional independence:

$$\mathbb{E}(Y_i \mid \bar{S}_{1:i}, \bar{X}^{\star}_{1:i}, \bar{X}_{1:i}; \mathcal{M}_\pi) = \mathbb{E}(Y_i \mid S_i, X^{\star}_i, X_i; \mathcal{M}_\pi). \tag{39}$$

These relations assert that in the counterfactual submodel, the future (state and intended action) and the reward depend only on the current tuple $(S_i, X^{\star}_i, X_i)$, and are conditionally independent of the full past history. Thus, despite the presence of unobserved confounding, the agent can reason with structured temporal dependencies within $\mathcal{M}_\pi$.

Lemma 2 (Invariance of ETTs across stages). Let $\mathcal{M}$ be an MDP environment, $\pi(X_i \mid S_i, X_i^{\textasciigrave})$ be a counterfactual policy over actions $X_1, \ldots, X_H$, and let $\mathcal{M}_\pi$ denote the induced counterfactual submodel. Then, for every decision stage $i = 1, \ldots, H$, the transition distribution over the next state S_{i+1} and next intended action X_{i+1}, as well as the expected reward Y_i, conditional on $(S_i, X_i^{\textasciigrave}, X_i)$ in the submodel $\mathcal{M}_\pi$, is equal to the corresponding *effect of treatment on the treated* (ETT) quantities in $\mathcal{M}$:

$$P(S_{i+1}, X_{i+1} \mid= s_i, x_i^{\textasciigrave}, x_i; \mathcal{M}_\pi) = P\big(S_{i+1}X_i{\leftarrow} x_i, X_{i+1}X_i{\leftarrow} x_i \mid s_i, x_i^{\textasciigrave}; \mathcal{M}\big), \tag{40}$$

$$\mathbb{E}(Y_i \mid S_i = s_i, X_i^{\textasciigrave} = x_i^{\textasciigrave}, X_i = x_i; \mathcal{M}_\pi) = \mathbb{E}\big[Y_i X_i{\leftarrow} x_i \mid s_i, X_i^{\textasciigrave} = x_i^{\textasciigrave}; \mathcal{M}\big]. \tag{41}$$

Moreover, these quantities remain invariant across every stage $i = 1, \ldots, H$, i.e.,

$$P\big(S_{i+1}X_i{\leftarrow} x_i, X_{i+1}X_i{\leftarrow} x_i \mid s_i, X_i^{\textasciigrave} = x_i^{\textasciigrave}\big) = \cdots = P\big(S_2X_1{\leftarrow} x_1, X_2X_1{\leftarrow} x_1 \mid s_1, X_1^{\textasciigrave} = x_1^{\textasciigrave}\big), \tag{42}$$

$$\mathbb{E}\big[Y_i X_i{\leftarrow} x_i \mid s_i, X_i^{\textasciigrave} = x_i^{\textasciigrave}\big] = \cdots = \mathbb{E}\big[Y_1 X_1{\leftarrow} x_1 \mid s_1, X_1^{\textasciigrave} = x_1^{\textasciigrave}\big]. \tag{43}$$

These invariances enable counterfactual generalization across stages, allowing policies optimized via counterfactual randomization to consistently exploit temporal structures in the environment.

Reduction to Standard MDP. Lemma 2 allows us to represent the joint distribution $P^\pi(S, X^{\textasciigrave}, X, Y)$ induced by a counterfactual policy

$$\pi = \big(\pi_1(X_1 \mid S_1, X_1^{\textasciigrave}), \ldots, \pi_H(X_H \mid S_H, X_H^{\textasciigrave})\big)$$

using a standard MDP of the form

$$\mathcal{M}_{\text{ctf}} = \big\langle \mathcal{D}(S), \mathcal{D}(X), \mathcal{D}(X), T_{\text{ctf}}, R_{\text{ctf}} \big\rangle,$$

where $\mathcal{D}(S)$ and $\mathcal{D}(X)$ denote the domains of state S_i and action X_i, respectively, at every stage $i = 1, \ldots, H$.

ETT-based Transition and Reward Functions. The transition distribution T_{ctf} and reward function R_{ctf} are defined in terms of conditional *effects of treatment on the treated* (ETTs) evaluated in the original MDP environment $\mathcal{M}$, as:

$$T_{\text{ctf}}(s, x, x^{\text{`}}, s', x') = P\left(S_{i+1}X_i{\leftarrow}\, x^{\text{`}} = s',\ X_{i+1}X_i{\leftarrow}\, x^{\text{`}} = x' \mid S_i = s,\ X_i = x\right), \tag{44}$$

$$R_{\text{ctf}}(s, x, x^{\text{`}}) = \mathbb{E}\left[Y_i X_i{\leftarrow}\, x^{\text{`}} \mid S_i = s,\ X_i = x\right]. \tag{45}$$

Planning Objective. Let $R(Y) = \sum_{i=1}^{\infty} \gamma^{i-1} Y_i$ be a discounted cumulative reward function over an infinite horizon $H = \infty$, with discount factor $\gamma \in (0, 1)$. The objective is to find an optimal counterfactual policy $\pi^\star \in \Pi_{\text{CTF}}$ that maximizes the expected reward

$$\pi^\star = \arg\max_{\pi \in \Pi_{\text{CTF}}} \mathbb{E}_\pi[R(Y)].$$

Stationary Policy Assumption. In light of classical planning results (Puterman, 1994), it is sufficient to consider a stationary class of counterfactual policies:

$$\pi = \left(\pi(X_i \mid S_i, X_i^{\text{`}})\right)_{i=1}^{\infty}, \quad \text{such that } \pi_1 = \pi_2 = \cdots.$$

Q-function and Augmented Bellman Equation. Define the Q-function for a counterfactual policy π as the expected discounted reward starting from state s, intended action x, and realized action $x^{\text{`}}$:

$$Q^\pi(s, x, x^{\text{`}}) = \mathbb{E}\left[\sum_{j=0}^{\infty} \gamma^j Y_{i+j} \mid S_i = s,\ X_i = x,\ X_i^{\text{`}} = x^{\text{`}};\ \mathcal{M}^\pi\right], \tag{46}$$

which can be recursively expressed using the following augmented Bellman equation:

$$Q^\pi(s, x, x^{\text{`}}) = R_{\text{ctf}}(s, x, x^{\text{`}}) + \gamma \sum_{s', x'} T_{\text{ctf}}(s, x, x^{\text{`}}, s', x') \sum_{x''} \pi(x'' \mid s', x') Q^\pi(s', x', x''). \tag{47}$$

Optimal Policy. An optimal counterfactual policy $\pi^\star$ can be obtained by iteratively computing Q^π and maximizing over realized actions $x^{\text{`}}$ for each (s, x) pair:

$$\pi^\star(x^{\text{`}} \mid s, x) = \arg\max_{x^{\text{`}}} Q^\pi(s, x, x^{\text{`}})$$

Relation to Interventional Policies. Experimental policies in the space

$$\Pi_{\text{EXP}} = \left\{\pi^{\text{`}} = (\pi^{\text{`}} i(X_i \mid S_i))_{i=1}^H\right\}$$

can be embedded in the counterfactual policy space

$$\Pi_{\text{CTF}} = \left\{\pi = (\pi_i(X_i \mid S_i, X^{\text{`}} i))_{i=1}^H\right\}.$$

For every experimental policy $\pi^{\text{`}}$, there exists a corresponding counterfactual policy π such that

$$\pi_i^{\text{`}}(x^{\text{`}} \mid s) = \pi_i(x^{\text{`}} \mid s, x), \quad \text{for all } s,\ x,\ x^{\text{`}}$$

In this case, the intended actions X do not influence the reward variables Y in the submodel $\mathcal{M}^\pi$ induced by the counterfactual intervention $\text{ctf}(\pi)$. Marginalizing out the intended actions X from the submodel $\mathcal{M}^\pi$ yields the experimental submodel $\mathcal{M}^{\pi^{\text{`}}}$ induced by the interventional action $\text{do}(\pi^{\text{`}})$.

As a result, the expected performance under both policies coincide:

$$\mathbb{E}_\pi[R(Y)] = \mathbb{E}_{\pi^{\text{`}}[R(Y)]}.$$

Theorem 9 (Counterfactual Dominates Interventional Policies)
Let $\mathcal{M}$ be an MDP environment, and let $R(Y)$ be a reward function over reward signals $Y = \{Y_1, \ldots, Y_H\}$. Let the counterfactual policy space be

$$\Pi_{\text{CTF}} = \left\{\pi = (\pi_i(X_i \mid S_i, X^{\text{`}} i))_{i=1}^H\right\}$$

and the interventional (experimental) policy space be

$$\Pi_{\text{EXP}} = \left\{\pi^{\text{`}} = (\pi^{\text{`}} i(X_i \mid S_i))_{i=1}^H\right\}.$$

Then,

$$\arg\max_{\pi \in \Pi_{\text{CTF}}} \mathbb{E}_\pi[R(Y)] \supseteq \arg\max_{\pi^{\text{`}} \in \Pi_{\text{EXP}}} \mathbb{E}_{\pi^{\text{`}}[R(Y)]}.$$

That is, the optimal performance of counterfactual policies weakly dominates that of interventional policies.

Corollary 3 (Equivalence under NUC)
Let $\mathcal{M}$ be an MDP environment in which the No Unobserved Confounding (NUC) condition holds. Then, for any decision stage $i = 1, \ldots, H$, state $s_i \in \mathcal{S}$, action $x_i \in \mathcal{X}$, and intended action $x^{\text{`}}i \in \mathcal{X}$,

$$\mathbb{P}\left(S^{i+1}_{X_i \leftarrow x^{\text{`}}i}, X^{i+1}_{X_i \leftarrow x^{\text{`}}i} \mid S_i = s_i, X_i = x_i\right) = \mathbb{P}_{X_i \leftarrow x^{\text{`}}i}(S_{i+1} \mid S_i = s_i), \tag{48}$$

$$\mathbb{E}\left[Y^{i}_{X_i \leftarrow x^{\text{`}}i} \mid S_i = s_i, X_i = x_i\right] = \mathbb{E}_{X_i \leftarrow x^{\text{`}}i}[Y_i \mid S_i = s_i]. \tag{49}$$

As a result, the counterfactual transition probabilities T_{ctf} and reward function R_{ctf} reduce to their experimental counterparts T_{exp} and R_{exp}. Hence, the performance of the optimal counterfactual policy coincides with that of the optimal experimental policy.

6.2 Counterfactual Randomization

In previous sections, we developed algorithms for computing optimal counterfactual policies that account for intended actions in environments such as MABs and MDPs. However, when the environment is unknown, a central question arises: how can one learn an optimal counterfactual policy by estimating counterfactual quantities from observed interactions?

To explore this, consider the MAB environment as an example. When the intended action produced by the agent's behavioral mechanism f_X matches the realized action (i.e., $x^{\text{`}} = x$), the counterfactual effect $\mathbb{E}[Y_x \mid x]$ aligns with the observed reward $\mathbb{E}[Y \mid X = x]$ due to the composition axiom [3, Ch. 7.3]. In contrast, when $x^{\text{`}} \neq x$, the intention-specific effect $\mathbb{E}[Y_x \mid x^{\text{`}}]$ is generally unidentifiable from passive data or standard interventions, unless further assumptions or full model knowledge are available.

Even in identifiable cases, implementing such a policy poses an additional challenge: agents often revise their intentions during deliberation. For example, an agent may initially consider $x^{\text{`}} = x_1$, later revise to $x^{\text{`}} = x_2$, and finally act on a different decision $x^{\text{`}} = x_t$. Only the last choice reflects the agent's final commitment, irrespective of how they reached it.

To operationalize counterfactual reasoning in such settings, we introduce a novel randomization method called *Counterfactual Randomized Controlled Trials* (Ctf-RCT). This procedure aims to intervene at the intention level: the agent's action deliberation is interrupted, the current intended action is treated as fixed, and the system proceeds with random or strategic exploration conditioned on that intention.

Algorithm 8: Counterfactual Randomized Controlled Trials (Ctf-RCT) in MAB

Require: *The domain of actions $\mathcal{D}(X)$; number of exploration episodes $N \in \mathbb{N}$.*

1. *For each episode $t = 1, 2, \ldots$:*
   (a) *Observe and store the agent's **intended** action $X^{(t)} \sim f_X$.*
   (b) *Select the **realized** action $X^{`}(t)$ as:*

$$X^{`}(t) = \begin{cases} \textit{Unif}(\mathcal{D}(X)) & \textit{if } t \leq N \\ \arg\max\limits_{x \in \mathcal{D}(X)} \widehat{\mathbb{E}}_N[Y_x \mid X = X^{(t)}] & \textit{if } t > N \end{cases}$$

   (c) *Apply the intervention $do(X^{`}(t))$ and observe the reward $Y^{(t)}$.*

Algorithm above outlines this process. Ctf-RCT takes as input the action space $\mathcal{X}$ and the number of exploration episodes N. For each episode t, the agent produces an intended action $X^{(t)} \sim f_X$. If $t \leq N$, the algorithm samples the realized action uniformly at random from $\mathcal{X}$. If $t > N$, the realized action is chosen to maximize the empirical estimate of the x-specific effect conditioned on the intended action $X = X^{(t)}$, using data collected during the exploration phase.

To estimate the counterfactual quantity $\mathbb{E}[Y_{X \leftarrow x} \mid X = x^{`}]$, we use the empirical mean over t episodes:

$$\hat{\mathbb{E}}^{(t)}[Y_{X \leftarrow x} \mid X = x^{`}] = \frac{1}{N_t(x^{`}, x)} \sum_{i=1}^{t} Y^{(i)} \cdot \mathbb{1}\{X^{(i)} = x,\, X^{`(i)} = x^{`}\}, \tag{50}$$

where $N_t(x^{`}, x)$ counts the number of times the intended action was $x^{`}$ and the executed action was x.

This quantity consistently estimates the effect of X on Y when intervening on $X^{`}$. The *Counterfactual Randomized Controlled Trial* (Ctf-RCT) chooses the intended action $X^{`}$ and intervenes with $do(X^{`})$ at each episode, observing the resulting reward.

To optimize decisions under uncertainty, this approach extends to a counterfactual version of UCB. For each pair $(x, x^{`})$, we define an upper confidence bound (UCB) using Hoeffding's inequality:

$$\text{UCB}_t(x, x^{`}, \delta) = \hat{\mathbb{E}}^{(t)}[Y_{X \leftarrow x^{`}} \mid X = x] + \sqrt{\frac{\log(1/\delta)}{2N_t(x, x^{`})}}, \tag{51}$$

where $\delta \in (0,1)$ is the error tolerance. At each episode t, the *Counterfactual UCB* (Ctf-UCB) algorithm observes the agent's intended action $X^{\backprime}(t)$, calculates UCBs over all x conditioned on $X^{\backprime}(t)$, and selects the action x maximizing $\mathrm{UCB}_t(x, X^{\backprime}(t), \delta)$. It then executes $do(X(t))$ and receives reward $Y^{(t)}$.

The value of δ is typically adapted across episodes, often set as a non-increasing function of the frequency $N_t(X^{\backprime}(t))$ of the intended action so far.

Algorithm 9: Counterfactual UCBVI in MDP (Ctf-UCBVI)

Input: *Counterfactual policy space* Π, *where* $\Pi = \{\backprime X_i, \{S_i, X_i\} \equiv\}_{i=1}^{H}$, *and a reward function* $R(Y) = \sum_{i=1}^{H} Y_i$.

1. *Initialize dataset* $\mathcal{H}^{(0)} = \emptyset$.
2. ***For*** *episodes* $t = 1, 2, \ldots$:
   (a) *Compute UCB Q-values:* $\hat{Q}^{(t-1)} \leftarrow \texttt{UCB-Q-values}(\mathcal{H}^{(t-1)})$.
   (b) ***For*** *step* $i = 1, \ldots, H$:
      i. *Observe current state* $S_i = S_i^{(t)}$.
      ii. *Intercept intended action* $X_i = X_i^{(t)}$.
      iii. *Choose counterfactual action:*

$$\backprime X_i^{(t)} = \arg\max_x \hat{Q}_i^{(t-1)}\left(S_i^{(t)}, X_i^{(t)}, x\right)$$

      iv. *Perform intervention* $\texttt{do}(X_i^{(t)} \leftarrow \backprime X_i^{(t)})$ *and receive reward* $Y_i^{(t)}$.
      v. *Update dataset:*

$$\mathcal{H}^{(t)} \leftarrow \mathcal{H}^{(t-1)} \cup \left\{S_i^{(t)}, X_i^{(t)}, \backprime X_i^{(t)}\right\}$$

Algorithm 10: UCB-Q-values

Input: *Dataset* $\mathcal{H}^{(t)} = \{S_i^{(k)}, X_i^{(k)}, \backprime X_i^{(k)}, Y_i^{(k)}\}_{k=1}^{t}$

1. *For all* $s, s' \in \mathcal{D}(S)$ *and* $x, x', x'' \in \mathcal{D}(X)$*, compute from* $\mathcal{H}^{(t)}$*:*

$$N_t(s, x, x') = \sum_{k=1}^{t} \sum_{i=1}^{H} \mathbb{1}\{S_i^{(k)} = s, X_i^{(k)} = x, 'X_i^{(k)} = x'\}$$

$$N_t'(s, x, x') = \sum_{k=1}^{t} \sum_{i=1}^{H} Y_i^{(k)} \cdot \mathbb{1}\{S_i^{(k)} = s, X_i^{(k)} = x, 'X_i^{(k)} = x'\}$$

$$N_t(s, x, x', s', x'') = \sum_{k=1}^{t} \sum_{i=1}^{H} \mathbb{1}\{S_i^{(k)} = s, X_i^{(k)} = x, \\ 'X_i^{(k)} = x', S_{i+1}^{(k)} = s', 'X_{i+1}^{(k)} = x''\}$$

2. *Let* $\mathcal{K} = \{(s, x, x') \in \mathcal{D}(S) \times \mathcal{D}(X) \times \mathcal{D}(X) \mid N_t(s, x, x') > 0\}$*.*

3. *For all* $(s, x, x') \in \mathcal{K}$*, compute:*

$$\hat{T}_t(s, x, x', s', x'') = \frac{N_t(s, x, x', s', x'')}{N_t(s, x, x')},$$

$$\hat{R}_t(s, x, x') = \frac{N_t'(s, x, x')}{N_t(s, x, x')}$$

4. *Initialize* $V_t^{(H+1)}(s, x) = 0$ *and* $Q_t^{(H)}(s, x, x') = H$ *for all* $(s, x, x') \in \mathcal{D}(S) \times \mathcal{D}(X) \times \mathcal{D}(X)$*.*

5. ***For*** $h = H, H-1, \ldots, 1$*:*

   (a) *For all* $(s, x, x') \in \mathcal{K}$*, compute:*

$$Q_t^{(h)}(s, x, x') = \min \Big\{ Q_t^{(h-1)}(s, x, x'), H, \hat{R}_t(s, x, x') + \hat{T}_t \\ \cdot V_t^{(h+1)}(s, x, x') + b_t^{(h)}(s, x, x') \Big\}$$

   where the bonus term is:

$$b_t^{(h)}(s, x, x') = 7H \sqrt{\frac{\log\left(5|\mathcal{D}(S) \times \mathcal{D}(X) \times \mathcal{D}(X)|T/\delta\right)}{N_t(s, x, x')}}$$

   (b) *Update value function:*

$$V_t^{(h)}(s, x) = \max_{x'} Q_t^{(h)}(s, x, x')$$

6.2.1 Counterfactual Randomization For MDPS

We extend the idea of counterfactual randomization to the setting of Markov Decision Processes (MDPs), enabling agents to reason over their intended actions and learn optimal policies in the counterfactual space ΠCTF.

Consider an unknown finite-horizon MDP $\mathcal{M}$ with horizon H, where the agent interacts over episodes $t = 1, 2, \ldots, T$. At each episode t, the agent selects a counterfactual policy:

$$\pi^{(t)} = \left(\pi_1^{(t)}(X_1 \mid S_1, X_1^{\text{`}}), \ldots, \pi_H^{(t)}(X_H \mid S_H, X_H^{\text{`}})\right),$$

where X_i is the action taken at step i, S_i is the current state, and $X_i^{\text{`}}$ is the agent's natural (intended) action at that step.

After executing $\pi^{(t)}$, the agent performs the intervention ctf-$\pi^{(t)}$ and observes a trajectory of rewards:

$$Y^{(t)} = \left(Y_1^{(t)}, \ldots, Y_H^{(t)}\right),$$

with the total reward defined as $R(Y^{(t)}) = \sum_{i=1}^{H} Y_i^{(t)}$.

Counterfactual Reward Function

We assume access to a known counterfactual reward function:

$$R_{\text{ctf}}(s, x, x^{\text{`}}) = \mathbb{E}[Y_i \mid S_i = s, X_i = x, X_i^{\uparrow} = x^{\text{`}}],$$

which allows evaluating expected rewards under hypothetical actions.

Ctf-UCBVI Algorithm

To solve the MDP under counterfactual reasoning, we adapt the UCBVI algorithm (Azar et al., 2017) into a counterfactual version called **Ctf-UCBVI**. The procedure iterates as follows:

1. At each episode t, based on historical data $\mathcal{H}^{(t-1)}$, compute the counterfactual Q-values $\hat{Q}_t(s, x, x^{\text{`}})$ using empirical Bellman updates and a confidence bonus derived via Chernoff-Hoeffding bounds.

2. Define the transition operator over values as:

$$\hat{\mathcal{T}}_t \cdot V_t^{(h+1)}(s, x, x^{\text{`}}) = \sum_{s', x'} \hat{T}_t(s, x, x^{\text{`}}, s', x') \cdot V_t^{(h+1)}(s', x'),$$

   where $\hat{T}_t$ estimates the counterfactual transition probabilities.

3. For each step $i = 1, \ldots, H$ of episode t:

   - Observe the current state $S_i = S_i^{(t)}$ and the agent's intended action $X_i^{`} = X_i^{(t)}$.
   - Choose an action $X_i^{(t)}$ by maximizing $\hat{Q}_t(S_i, x, X_i^{`})$ over x.
   - Execute the intervention $do(X_i^{(t)})$ and observe reward $Y_i^{(t)}$.

7 Causal Imitation Learning

Reinforcement Learning (RL) has been used successfully in very complex environments over the past few decades. A key assumption in many traditional RL algorithms is that the reward function is clearly defined. However, in many real world situations, it is difficult to create a reward function that covers every possible case. For instance, in human driving, it's hard to design an accurate reward function, and testing in the real world can be risky. But observing expert drivers is usually possible. In reinforcement learning, the imitation learning (IL) approach looks at how an agent can learn to act in an environment where the reward function is unknown, by watching demonstrations from a human expert. Formally, a causal imitation learning task is described by the following Causal Decision Model.

$$\mathcal{T} = \langle \mathcal{I} = \text{see},\ \mathcal{A} = \mathcal{G},\ \Pi = \{\langle \mathbf{X}_i,\ \mathbf{S}_i \rangle\}_{i=1}^{H},\ \mathcal{R} = \phi \rangle \tag{52}$$

The Agent will try to find a policy such that,

$$\pi^* = \arg \max_{\pi \in \Pi_{\text{EXP}}} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R} \,|\, \mathcal{G},\ \mathcal{D}_{\text{obs}} \sim \mathbf{P}(\mathbf{V}),\ \mathcal{R} = \phi \right] \tag{53}$$

An imitation learning agent aims to acquire a policy within a space by leveraging demonstration data provided by a demonstrator (such as a driver or physician) who follows a distinct behavioral policy. In each episode $t = 1, \ldots, T$, the agent observes the demonstrator interacting with the environment $\mathcal{M}^*$ and receives trajectories $V^{(t)} \sim P(V)$ sampled from the corresponding observational distribution. The agent may also have access to structural knowledge about the environment, such as a causal graph $\mathcal{A} = G$. Unlike off policy learning or causal inference tasks, imitation learning is uniquely challenging because the reward function is neither disclosed to the agent nor explicitly defined (i.e., $R = \varnothing$).

Corollary 4
Let $X, Y \subseteq V$ be endogenous variables. If the reward function $R : \mathcal{D}(Y) \to \mathbb{R}$ is not explicitly specified, then the expected reward $\mathbb{E}_{\pi}[R(Y)]$ under any policy π over actions X cannot be identified from the causal diagram G.

As stated in Corollary 4, when the reward function R is unknown, the expected reward $\mathbb{E}_{\pi}[R(Y)]$ under a policy π cannot be uniquely determined from the observational distribution $P(V)$ within the causal diagram G.

To address this challenge of non identifiability, a common strategy is to assume that the observed trajectories are generated by an expert demonstrator whose performance, measured by $\mathbb{E}[R(Y)]$, meets or exceeds a known threshold. If we can identify a policy π that achieves at least the expert's level of performance, we can ensure that the agent's policy is also satisfactory.

Definition 23
For a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$, an imitating policy π^* is a policy such that its expected reward is lower bounded by the expert's reward, i.e.,

$$\underbrace{\mathbb{E}_{\pi^*}\left[\mathcal{R}(\mathbf{Y}); \mathcal{M}^*\right]}_{\textit{Agent's Performance}} \geq \underbrace{\mathbb{E}\left[\mathcal{R}(\mathbf{Y}); \mathcal{M}^*\right]}_{\textit{Expert's performance}}$$

The core principle of imitation learning is captured by a fundamental inequality: the agent's expected reward should match or exceed that of the expert when evaluated in the underlying environment $\mathcal{M}^*$. On the left hand side of this equation lies the agent's actual performance, while the right hand side denotes the expert's performance; serving as the benchmark the agent aims to achieve. The ultimate goal in imitation learning is to find a policy π^* that satisfies this condition.

Two main paradigms dominate the landscape of imitation learning:

- **Behavioral Cloning (BC)**: These methods aim to directly replicate the expert's policy by learning a mapping from states to actions using supervised learning.
- **Inverse Reinforcement Learning (IRL)**: These approaches first infer a reward function under which the expert's policy is optimal, and then derive an optimal policy for the agent by maximizing this learned reward.

Under standard assumptions, both BC and IRL can produce agent policies that achieve performance on par with the expert. In some cases, especially when additional structural information about the reward function is available, IRL can even yield policies that outperform the expert in the same environment.

However, the effectiveness of both BC and IRL hinges on a critical assumption: that the agent has access to the same observations as the expert. When some of the expert's observed variables are hidden from the agent, the

demonstrations are influenced by unobserved confounders (UCs), violating the no unobserved confounding (NUC) assumption.

In such cases, applying BC or IRL directly may fail to yield satisfactory results, even if the expert's behavior is optimal. This challenge raises a fundamental question: how can one perform imitation learning in the presence of unobserved confounding? The remainder of this section addresses this question and explores imitation learning from a causal perspective.

7.1 Causal Behavioral Cloning

We begin with the *behavioral cloning* (BC) approach, where the agent seeks to learn an imitating policy by directly replicating the conditional observational distributions $P(X_i \mid Z_i)$. Here, each action X_i is predicted based on a subset of input states $Z_i \subseteq S_i$, where S_i denotes the full set of variables observed at step i.

Definition 24 (Behavioral Cloning Policy)
Let $\Pi = \{\Pi_i(X_i \mid Z_i)\}_{i=1}^{H}$ denote a policy space, and let $P(V)$ represent an observational distribution over variables V. A behavioral cloning policy $\pi \in \Pi$ is defined in terms of $P(V)$ such that, for each time step $i = 1, \ldots, H$, we have

$$\pi_i(X_i \mid Z_i) = P(X_i \mid Z_i) \quad \textit{for some } Z_i \subseteq S_i.$$

Example 3 (Behavioral Cloning in a Multi Armed Bandit Setting)
Consider a simple environment where an agent must learn a driving policy that imitates the behavior of a car following another. The policy $\pi(X)$ belongs to a space where actions depend on a variable U, such as the distance to the front car.

Using the behavioral cloning approach, the agent directly imitates the observed action distribution. Specifically, the behavioral cloning policy $\pi_{BC}(X)$ is defined as:

$$\pi_{BC}(X = 1) = P(X = 1) = P(U = 1)$$

Assuming $P(U = 1) = 0.9$, this yields $\pi_{BC}(X = 1) = 0.9$.

Suppose the reward is defined as a linear function of some outcome Y, written as $\mathcal{R}(Y) = \varepsilon Y$, where ε is an unknown constant. The expected reward under the behavioral cloning policy is then:

$$\mathbb{E}_{\pi_{BC}}[\mathcal{R}(Y)] = \varepsilon \cdot \pi_{BC}(X = 1) = 0.9\varepsilon$$

To compare this with the expert's performance, consider the expected reward under the expert's true behavior. Since the expert acts based on U, and

$P(U = 1) = 0.9$, *we have:*

$$\mathbb{E}[\mathcal{R}(Y)] = \varepsilon \cdot \mathbb{E}[Y] = \varepsilon \cdot P(U = 1) = 0.9\varepsilon$$

Thus, in this scenario, the behavioral cloning policy successfully replicates the expert's expected performance.

Theorem 10 (Behavioral Cloning under No Unobserved Confounding)
Let $\langle\ \mathcal{M},\ \Pi = \{\pi_i(X_i \mid S_i)\}_{i=1}^{H},\ \mathcal{R} : Y \mapsto \mathbb{R}\ \rangle$ *define a causal decision model (CDM), where* Π *is the space of policies over actions* X_i *conditioned on subsets of observed variables* S_i*.*

Suppose the following conditions hold:

1. *The no unobserved confounding (NUC) condition is satisfied for the policy space* Π *in the structural causal model* $\mathcal{M}$*;*

2. *For each action* X_i *at time step* $i = 1, \ldots, H$*, its endogenous parents in the model are a subset of the observed variables:* $PA_i \subseteq S_i$*.*

Then, there exists a behavioral cloning policy $\pi \in \Pi$ *such that the expected reward under this policy matches the expert's performance in the true environment* $\mathcal{M}$*:*

$$\mathbb{E}_{\pi}[\mathcal{R}(Y); \mathcal{M}] = \mathbb{E}[\mathcal{R}(Y); \mathcal{M}]$$

Moreover, such a policy is explicitly given by $\pi_i(X_i \mid S_i) = P(X_i \mid S_i)$ *for all* $i = 1, \ldots, H$*.*

Example 4 (Failure of Behavioral Cloning without NUC)
Revisit the driving scenario where the agent aims to learn a policy by observing a human driver. In this setup, the distance Y *between the demonstrator and the front car depends on the deceleration* U *of the front car. The human driver observes* U *through the tail light and uses it to decide on an action* X*. However, the imitator only has access to drone footage, in which the tail light (i.e.,* U*) is not visible. Thus,* U *remains unobserved from the imitator's perspective.*

The environment can be modeled by a structural causal model $\mathcal{M}$ *defined over:*

$$\mathcal{M} = (\mathbf{U} = \{U\}, \mathbf{V} = \{X, Y\}, F, P(U))$$

where the structural equations are:

$$X \leftarrow \neg U$$
$$Y \leftarrow X \oplus U$$

Here, $U \in \{0, 1\}$ is a binary variable drawn from $P(U = 1) = 0.5$. Since U influences both the action X and the outcome Y, and is unobserved by the imitator, the no unobserved confounding (NUC) condition does not hold in this environment.

Let the driver's performance be measured by a reward function $\mathcal{R}(Y) = \varepsilon Y$, where $\varepsilon \in \mathbb{R}$ is a fixed but unknown parameter. The expected reward under the expert's policy is:

$$\mathbb{E}[\mathcal{R}(Y)] = \varepsilon \cdot \mathbb{E}[Y] = \varepsilon \cdot \mathbb{E}[X \oplus U] = \varepsilon \cdot \mathbb{E}[\neg U \oplus U] = \varepsilon$$

Now, applying the behavioral cloning approach, the imitator mimics the marginal distribution of the expert's actions:

$$\pi_{BC}(X = 1) = P(X = 1) = P(U = 1) = 0.5$$

This results in a stochastic policy where the agent chooses to accelerate half the time, independent of the unobserved U.

To evaluate the performance of this policy, consider the expected reward under the new induced model $\mathcal{M}_{\pi_{BC}}$:

$$\mathbb{E}_{\pi_{BC}}[\mathcal{R}(Y)] = \sum_x \varepsilon \cdot \mathbb{E}[Y \mid X = x] \cdot \pi_{BC}(x) = 0.5 \cdot \varepsilon \cdot (\mathbb{E}[Y \mid X = 0] + \mathbb{E}[Y \mid X = 1])$$

Since $Y = X \oplus U$, we compute:

$$\mathbb{E}_{\pi_{BC}}[\mathcal{R}(Y)] = 0.5 \cdot \varepsilon \cdot (\mathbb{E}[0 \oplus U] + \mathbb{E}[1 \oplus U]) = 0.5 \cdot \varepsilon$$

Hence, the performance under the behavioral cloning policy is:

$$\mathbb{E}_{\pi_{BC}}[\mathcal{R}(Y)] = 0.5\varepsilon$$

This is significantly worse than the expert's performance ε, highlighting that when the NUC condition is violated due to unobserved confounders, behavioral cloning fails to match expert performance.

7.1.1 Backdoor Criterion for Imitation

We begin our discussion with a modified version of the sequential backdoor criterion (Definition 15), which enables a learner to replicate the expert's behavior. Recall that for any policy $\pi \in \Pi$, the graph $G_{\pi_{i+1},\ldots,\pi_H}$, where $i = 0, \ldots, H-1$, denotes a modified causal graph derived from the original diagram G by replacing all incoming edges to each action node X_j (for $j \in \{i+1, \ldots, H\}$) with edges originating from the respective state inputs S_j to X_j. In imitation learning, $G_{\pi_{i+1},\ldots,\pi_H}$ represents the causal structure where all future actions beyond the i-th stage are embedded into the graph.

Definition 25 (Imitation Backdoor Condition)
Let G be a causal graph, and let $X, Y \subseteq V$ be variable subsets. A policy class $\Pi = \{X_i, S_i\}_{i=1}^{H}$ is said to satisfy the imitation backdoor condition with respect to Y in G (or simply, Π is imitation admissible*) if for every policy $\pi \in \Pi$ and each action node $X_i \in X$, one of the following holds:*

1. *X_i is not an ancestor of Y in the graph $G_{\pi_{i+1},\ldots,\pi_H}$, i.e., $X_i \notin An(Y)_{G^{\pi_{i+1},\ldots,\pi_H}}$;*
2. *The set S_i d-separates all backdoor paths from X_i to any node in Y in $G^{\pi_{i+1},\ldots,\pi_H}$, that is, $(Y \perp\!\!\!\perp X_i \mid S_i)_{G^{X_i,\pi_{i+1},\ldots,\pi_H}}$.*

The first condition in Definition 25 captures scenarios in which the action at X_i has no effect on the outcome Y once subsequent actions have been fixed. In the graph $G_{\pi_{i+1},\ldots,\pi_H}$, the parental structure of future actions $\bar{X}_{i+1:H}$ has been altered, potentially rendering the value of X_i irrelevant to Y. That is, the distribution over Y remains unchanged regardless of the policy selected for X_i. This permits X_i to fail Condition (2) indicating it cannot be imitated in isolation—yet still belong to a clonable set of actions X, since downstream actions may buffer Y from inaccuracies at X_i.

The second condition is analogous to the classical backdoor criterion, where Z_i is a variable set encoding sufficient information to support imitation of X_i with respect to Y. In practical terms, if the joint distribution $P(Z_i, X_i)$ over observed states Z_i and actions X_i aligns for both the expert and the imitator, then a reward signal Y generated adversarially cannot discern between the two. Consequently, effective imitation can be achieved.

The imitation backdoor condition provides a practical criterion for determining whether a policy, compatible with a given policy space in a causal diagram, can be effectively imitated. We now introduce several key notations required for this characterization.

Definition 26 (Policy Subspace)
Given a policy space $\Pi = \{X_i, S_i\}_{i=1}^{H}$, a policy subspace Π' *of Π, denoted by $\Pi' \subseteq \Pi$, is defined as a sequence $\{X_i, Z_i\}_{i=1}^{H}$ where $Z_i \subseteq S_i$ for each action $X_i \in X$.*

In other words, every policy space Π is trivially a subspace of itself. A subspace Π' is called proper *if $\Pi' \neq \Pi$, which we denote by $\Pi' \subsetneq \Pi$. For any policy $\pi' \in \Pi'$, one can always construct a corresponding policy $\pi \in \Pi$ such that $\pi_i(x_i \mid s_i) = \pi'_i(x_i \mid z_i)$ for all $i = 1, \ldots, H$ and for all realizations x_i, s_i, and z_i. This reflects that inputs in $S_i \setminus Z_i$ have no influence on the chosen action X_i. Therefore, a policy π' is also compatible with the larger space Π if it is compatible with a subspace $\Pi' \subseteq \Pi$.*

Theorem 11 (Behavioral Cloning from Imitation Backdoor)
Let G be a causal diagram, Π a policy space over action variables X, and let $Y \subseteq V$ be a set of variables. If there exists a subspace $\Pi^{'} = \{X_i, Z_i\}_{i=1}^{H}$ of Π such that $\Pi^{'}$ satisfies the imitation backdoor condition with respect to Y in G, then there exists a behavioral cloning policy $\pi \in \Pi^{'}$ such that for any structural causal model $\mathcal{M}$ compatible with G,

$$\mathbb{E}_{\pi}[R(Y); \mathcal{M}] = \mathbb{E}[R(Y); \mathcal{M}].$$

Moreover, such a policy is explicitly given by $\pi_i(x_i \mid z_i) = P(x_i \mid z_i)$ for each $i = 1, \ldots, H$.

Theorem 11 establishes that whenever an imitation admissible subspace $\Pi^{'} \subseteq \Pi$ is identified, the expert's performance can be replicated via behavioral cloning—specifically, by matching the conditional distribution $P(X_i \mid Z_i)$ for each action $X_i \in X$. Furthermore, it has been demonstrated that the imitation backdoor criterion is not only sufficient but also necessary for behavioral cloning to succeed within a broad class of policy spaces. In this class, each input state S_i contains all variables that temporally precede action X_i in the causal diagram G .

In particular, if no imitation admissible subspace exists within Π, then for any behavioral cloning policy $\pi \in \Pi$, one can always construct a structural causal model $\mathcal{M}$ consistent with G such that the resulting BC policy π fails to match the expert's performance.

7.2 Causal Inverse RL

This section explores an alternative approach to imitation learning, known as inverse reinforcement learning (IRL) , within a causal decision-making model $(\mathcal{M}, \Pi, R)$. As in behavioral cloning, the imitator does not possess full knowledge of the environment $\mathcal{M}$ or the true reward function R. However, the agent has access to a parametric family of reward functions $\mathcal{R} = \{\hat{R} : D(Y) \to \mathbb{R}\}$ that contains the actual reward R.

We begin by analyzing an IRL strategy under the assumption of No Unmeasured Confounding (NUC; Definition 13), and subsequently relax this assumption to accommodate a broader class of causal models specified by a causal diagram G.

Following the game-theoretic formulation of , imitation learning (Definition 28) is modeled as a two-player zero-sum game. In this game, the agent selects a policy, while an adversary (e.g., Nature) chooses a worst-case reward function from the set $\mathcal{R}$. The learning objective is expressed as the following min-max optimization:

$$\Delta = \min_{\pi \in \Pi} \max_{R \in \mathcal{R}} \left(\mathbb{E}_{\pi^*}[R(Y); \mathcal{M}] - \mathbb{E}_{\pi}[R(Y); \mathcal{M}] \right).$$

Here, the inner maximization can be interpreted as a causal IRL step that identifies a worst-case reward function $\hat{R} \in \mathcal{R}$ which maximizes the discrepancy in performance between the expert policy π^* and a candidate imitation policy π. Notably, the expert's expected reward $\mathbb{E}_{\pi^*}[R(Y); \mathcal{M}]$ is independent of the choice of π.

The outer minimization corresponds to planning within a causal decision-making model $(\mathcal{M}, \Pi, \hat{R})$ using the adversarial reward $\hat{R}$. If the performance gap $\Delta = 0$, then the solution $\pi^\star$ is a successful imitation of the expert. If the expert policy is sub-optimal, it is possible to observe $\Delta < 0$, i.e.,

$$\mathbb{E}_{\pi^*}[\hat{R}(Y); \mathcal{M}] < \mathbb{E}_{\pi^\star}[\hat{R}(Y); \mathcal{M}],$$

indicating that the learned policy $\pi^\star$ dominates the expert in the worst-case scenario. This suggests that $\pi^\star$ may choose to disregard the potentially flawed expert policy, leveraging prior structural knowledge about the reward function. When this prior is informative, solving the optimization problem above can yield a policy that not only matches but potentially exceeds the expert's performance in the true environment—despite the reward function remaining partially unknown.

Despite the clear semantics of the objective, solving the associated optimization problem requires detailed knowledge of the underlying structural causal model $\mathcal{M}$, which is generally unavailable to the agent in practical scenarios. This raises the question of identifiability: under what conditions can the solution to Eq. 474 be inferred from the observational distribution $P(V)$?

Let us fix a reward function $R \in \mathcal{R}$. The expert's expected reward $\mathbb{E}_{\pi^*}[R(Y); \mathcal{M}]$ can be computed directly from $P(V)$ by evaluating the average of $R(Y)$ weighted by the marginal distribution $P(Y)$. Under the No Unmeasured Confounding (NUC) assumption, the imitator's expected reward $\mathbb{E}_{\pi}[R(Y); \mathcal{M}]$ is also estimable from $P(V)$ via standard off-policy evaluation methods, such as Inverse Propensity Weighting or Direct Policy estimation.

Consequently, under the NUC assumption, the agent can formulate the minimax optimization, using only the observational distribution $P(V)$, the hypothesis class $\mathcal{R}$ of reward functions, and the policy space Π. Solving this identifiable program yields a valid imitation policy.

7.2.1 Minimal Imitation Backdoor

We now turn to the study of causal inverse reinforcement learning (IRL) in more general settings where the No Unmeasured Confounding (NUC) assumption does not hold. In such scenarios, demonstration data may be influenced by latent confounders that affect both the agent's actions and other variables in the environment. Our approach builds upon a refinement of the imitation backdoor criterion (Definition 30), leveraging the concept of minimal d-separating sets.

Definition 27 (Minimal Imitation Backdoor)
Let G be a causal diagram, and let $X, Y \subseteq V$ denote sets of variables. An imitation admissible policy space Π over X is said to be minimal *if there exists no proper subspace $\Pi' \subsetneq \Pi$ that also satisfies the imitation backdoor criterion with respect to Y in G.*

In other words, a policy space $\Pi = \{X_i, S_i\}_{X_i \in X}$ is minimal if, for each action variable $X_i \in X$, the set S_i is a minimal d-separating set between X_i and the reward variable(s) Y in the manipulated causal graph $G_{X_i, \pi_{i+1}, \ldots, \pi_H}$. Alternatively, if X_i is not an ancestor of Y in the graph $G_{\pi_{i+1}, \ldots, \pi_H}$, then $S_i = \emptyset$.

Theorem 12 (Identifiability under Minimal Imitation Backdoor)
Let G be a causal diagram, let Π be a policy space over actions X, and let $Y \subseteq V$ be a subset of variables. Suppose there exists a subspace $\Pi' = \{X_i, Z_i\}_{i=1}^H$ contained within Π such that Π' is minimal imitation admissible with respect to Y in G. Then, for any policy π compatible with Π', the interventional distribution $P^\pi(Y)$ is identifiable from the observational distribution $P(V)$ and is given by:

$$P^\pi(y) = \sum_{\bar{x}_H, \bar{z}_H} P(y \mid \bar{x}_H, \bar{z}_H) \prod_{i=1}^{H} P(z_i \mid \bar{x}_{i-1}, \bar{z}_{i-1})\, \pi_i(x_i \mid z_i),$$

where $\bar{X}_i = \{X_1, \ldots, X_i\}$ and $\bar{Z}_i = \{Z_1, \ldots, Z_i\}$ denote the sequences of actions and covariates up to stage i, for $i = 1, \ldots, H$.

The notion of a minimal imitation backdoor introduced in Definition 32, together with the identifiability result established in Theorem 21, naturally leads to an algorithmic strategy for causal inverse reinforcement learning (IRL) in the presence of unobserved confounding. Rather than searching for imitating policies directly within the full policy space Π, the agent instead restricts attention to a minimal imitation admissible subspace $\Pi' \subseteq \Pi$.

As previously discussed in Section 8.1, there exist efficient procedures for identifying admissible policy subspaces that satisfy the imitation backdoor criterion. Given such an admissible subspace, one can refine it into a minimal admissible subspace by progressively eliminating input state variables from S_i for each decision point X_i, stopping once the imitation backdoor condition no longer holds. This reduction process can be executed in polynomial time with respect to the total number of state variables S and action variables X.

7.2.2 Imitation Via Inverse RL

7.3 Imitation via Inverse Reinforcement Learning

Once a minimal imitation admissible subspace $\Pi^{\text{`}} \subseteq \Pi$ is identified, an imitating policy can be derived by solving the minimax optimization problem in Eq. (474), where the original policy space $\Pi$ is replaced by the refined subspace $\Pi^{\text{`}}$. Expanding the objective with respect to the possible outcomes of Y, the optimization problem can be reformulated as follows:

$$\Delta = \min_{\pi \in \Pi^{\text{`}}} \max_{R \in \mathcal{R}} \sum_{y} R(y) \left(P(y) \quad \underbrace{-P_{\pi}(y)}_{\text{Performance gap}} \right) \tag{54}$$

In this formulation, the term $P(y)$ represents the expert's occupancy measure over the domain of reward-relevant variables Y, corresponding to the marginal distribution derived from observational data. The second term, $P_{\pi}(y)$, denotes the imitator's occupancy measure, expressed as an interventional distribution. Due to the minimal admissibility of $\Pi^{\text{`}}$, Theorem 21 ensures that $P_{\pi}(Y)$ is identifiable from the observational distribution $P(V)$ and the given policy π.

When the hypothesis class $\mathcal{R}$ for the reward functions is chosen appropriately, the minimax program in Eq. (54) becomes solvable via several well-established IRL techniques. Consequently, we refer to Eq. (54) as the *canonical IRL formulation*. In the following, we illustrate how this reduction enables the application of existing algorithms such as the multiplicative-weights algorithm (MWAL) and generative adversarial imitation learning (GAIL).

Causal MWAL

Causal MWAL investigates inverse reinforcement learning in Markov decision processes where the reward function $R(y)$ is modeled as a linear combination of feature expectations. Let the feature mapping be denoted by $\pi(y) \in \mathbb{R}^k$,

and assume that the reward function takes the form $R(y) = w \cdot \pi(y)$, where $w \in \mathcal{P}_k$ is a coefficient vector in a convex set defined as:

$$\mathcal{P}_k = \left\{ w \in \mathbb{R}^k \mid Aw \leq 1,\, w \geq 0 \right\} \tag{55}$$

Let $\pi^{(i)}$ denote the i-th component of the feature vector π. Suppose that deterministic policies within the space Π are indexed according to feature orderings $\pi^{(1)}, \ldots, \pi^{(n)}$. Then, the canonical IRL program in Eq. (54) reduces to a two-player zero-sum matrix game under the linearity assumption.

Proposition 4
Let the hypothesis class of reward functions be $\mathcal{R} = \{R = w \cdot \pi \mid w \in \mathcal{P}_k\}$. *Then the solution* Δ *of the canonical IRL program in Eq.* (54) *is equivalent to the solution of the following minimax problem:*

$$\Delta = \min_{\pi \in \Pi^{`}} \max_{w \in \mathcal{P}_k} w^\top G_\pi \tag{56}$$

where $G \in \mathbb{R}^{k \times n}$ *is a matrix with entries defined as*

$$G(i,j) = \sum_y \pi^{(i)}(y) \left(P(y) - P_{\pi^{(j)}}(y) \right)$$

Effective algorithms for solving the matrix game in Eq. (56) include the classical Multiplicative Weights (MW) method and its adaptation for apprenticeship learning, MWAL.

Causal GAIL, introduced by Ho and Ermon (2016), presents an approach for learning imitation policies in Markov Decision Processes using a flexible class of non-linear reward functions. The reward function $R(y)$ maps outcomes to real values, formally $R : \mathcal{Y} \to \mathbb{R}$, where $\mathcal{R}_\mathcal{Y} = \{r : \mathcal{D}(\mathcal{Y}) \to \mathbb{R}\}$.

To manage the complexity of the reward function, a convex regularization function $\Psi(R)$ is applied. The resulting optimization problem, known as the *penalized canonical program of causal IRL*, is expressed as:

$$\min_\pi \max_{R \in \mathcal{R}_\mathcal{Y}} \left[\mathbb{E}_{y \sim P}[R(y)] - \mathbb{E}_{y \sim P_\pi}[R(y)] - \Psi(R) \right]$$

This formulation is commonly referred to as Equation (502) in related literature. It is often more convenient to work with its dual form. The following proposition states this formally:

Theorem 13 (Proposition 7)
Given a hypothesis class $\mathcal{R} = \{R : \mathcal{D}(\mathcal{Y}) \to \mathbb{R}\}$ *regularized by* Ψ, *the solution* π *to the canonical program above can also be obtained by solving:*

$$\min_\pi \Psi^*(P \| P_\pi)$$

where Ψ^* is the convex conjugate of Ψ, defined as
$\Psi^*(P\|P_\pi) = \max_{R\in\mathcal{R}_\mathcal{Y}} [\mathbb{E}_{y\sim P}[R(y)] - \mathbb{E}_{y\sim P_\pi}[R(y)] - \Psi(R)]$.

Equation (503) aims to find a policy π that minimizes the divergence between the expert's distribution and the one induced by the policy over the outcome space $\mathcal{Y}$, with the divergence quantified by Ψ^*.

When a specific regularizer $\Psi(R)$, as described in Equation (13) of Ho and Ermon (2016), is used, the optimization problem in Equation (503) simplifies to:

$$\min_\pi \Psi^*(P\|P_\pi) = \min_\pi \max_{D:\mathcal{Y}\to(0,1)} [\mathbb{E}_{y\sim P}[\log D(y)] + \mathbb{E}_{y\sim P_\pi}[\log(1 - D(y))]]$$

Here, D denotes a discriminator function (often parameterized as a neural network) that estimates whether a sample comes from the expert or the policy. This formulation aligns causal imitation learning with the adversarial training paradigm of Generative Adversarial Networks (GANs), where two neural networks — a generator (policy π) and a discriminator — are trained in a minimax setup.

The policy π is considered successful when the discriminator D can no longer distinguish between the expert and the learner's trajectories. Solving the minimax problem involves finding a saddle point (π, D), which can be approached through iterative optimization methods as implemented in the GAIL algorithm by Ho and Ermon.

8 Conclusion

The current generation of AI agents that excel at decision-making are predominantly built on the foundations of reinforcement learning (RL). However, most RL systems lack explicit modeling of underlying causal structures and do not engage in causal reasoning.

Across diverse fields, there is a growing recognition that effective decision-making requires understanding causal relationships in the environment. For example, a robot must interpret causal dynamics in its surroundings to plan actions, a physician must know the effects of medications to treat patients effectively, and an economist must understand the causal link between education and employment to shape policy. These scenarios illustrate that decision-making often depends on grasping complex and sometimes hidden causal mechanisms.

Despite some progress toward integrating causal reasoning into RL, a systematic and cohesive framework has been largely missing. To address this, we integrate Pearl's Structural Causal Models (SCMs) with RL, enabling agents to encode causal knowledge and perform counterfactual reasoning. This fusion gives rise to a theoretical and algorithmic foundation for *Causal Reinforcement Learning (CRL)*, aimed at robust decision-making under uncertainty.

The CRL framework enhances RL algorithms in several ways. It allows us to relax key assumptions about the data-generating process and enables learning in the presence of unobserved confounders. We propose new algorithms for offline learning—such as off-policy and imitation learning—that remain reliable despite confounding. We also design online algorithms capable of identifying optimal policies while maintaining low regret, using causal insights even from biased offline data.

Finally, we distinguish between the different modes of agent interaction with the environment. While supervised learning relies on passive observation, RL agents learn by actively exploring and adapting. By generalizing this interaction—incorporating both passive and active elements—we expand the space of learning possibilities. In particular, *counterfactual randomization* extends classical randomized experiments into a counterfactual domain, empowering agents with the ability to reason about hypothetical scenarios. We argue this capacity is crucial for the development of the next generation of intelligent systems.

References

[1] Elias Bareinboim, Junzhe Zhang, and Sanghack Lee. Towards causal reinforcement learning. Technical Report R-65, Causal Artificial Intelligence Lab, Columbia University, December 2024.

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

[3] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.