Trends, Problems, and Solutions in Causality and Reinforcement Learning

St John M.M. Grimbly

Prepared in fulfilment of the requirements for the award of a Master of Science (MSc) degree in Applied Mathematics (MAM 5001W).

Supervised By
Assoc. Prof. Jonathan Shock

University of Cape Town, 2023
Department of Mathematics and Applied Mathematics

Abstract

This thesis reviews, examines, and investigates the trends in the fields of causality and in reinforcement learning (RL). Theory is developed for both active research areas, with a specific focus on the overlap in underlying theory. The core argument is that the RL problem can be formulated as a causal problem, where the agent is learning causal policies that maximise return (via some causal relationship implied by the policy) and does this via selecting optimal actions (performing interventions) in the environment. Although relevant in both model-based and model-free scenarios, focus is placed on model-based modalities where one can view the various models as being causal models. It is further argued that this reformulation enables various theoretical improvements in reasoning ability for a learning agent, and does this while offering improved efficiency, interpretability, robustness, and generalisation across various learning modalities. As an application of the causal methods discussed, we also investigate whether applied causal discovery can lead to disparate impacts on sensitive subgroups. Finally, we reflect on the findings, highlight open problems, and propose future research directions.

Acknowledgements

I am deeply indebted to Assoc. Prof. Jonathan Shock for his guidance and belief in my capabilities throughout this project. His constant encouragement and insightful discussions, especially regarding my nonlinear journey, have been important in shaping this thesis.

My heartfelt thanks go to my lab mates, especially Rachel Catzel, Roland Dubb, Jeremy du Plessis, and Claude Formanek. Our collaborative problem-solving sessions and discussions have made this process deeply enjoyable.

Special appreciation is extended to InstaDeep for their support in my first year of research. The opportunity to engage with practical problems in a collaborative environment laid a solid foundation for my research interests. I am particularly grateful to Dr Arnu Pretorius for his mentorship and guidance during this phase.

I am thankful for the enriching experience during my internship at Prof. Ferdinando Fioretto's lab in 2023. The discussions on causal fairness and optimisation, along with his insights on pursuing a PhD, have been incredibly enlightening.

I acknowledge with gratitude the funding from PGFO and ETDP-SETA during 2021 and 2022, which was instrumental in facilitating this research.

My gratitude is also extended to the University of Cape Town and the Department of Mathematics and Applied Mathematics. Their conducive and flexible academic environment played a crucial role in the completion of this project.

The unwavering support from my family and friends cannot be discounted. Their belief in my goals and abilities has made this all possible.

Finally, a big thank you to everyone who contributed, directly or indirectly, to this thesis. It has been a challenging yet immensely rewarding ride, and each one of you acknowledged here has been a significant part of making it a success. Your roles in this academic venture have been truly appreciated.

Ad Astra

Contents

List of Figures

List of Tables

List of Algorithms

List of Code Snippets

Notation

Throughout this paper we attempt to be consistent with notation. We specifically aim to do this across sections stemming from different fields, but this is not always achievable nor desirable. For example, when referring to standard literature in reinforcement learning, the set (size n) of possible actions will usually be represented as $\boldsymbol{A} = \{a_1, a_2, \ldots, a_n\}$. This differs from causality literature where interventions are represented by various letters, perhaps $\boldsymbol{I}$ (intervention) or $\boldsymbol{T}$ (treatment). In this case, we will choose what we deem to make the most sense in the context of the chapter or example. We now discuss the theme of how we notate in this paper.

Bold symbols (e.g. $\boldsymbol{X}$) will generally notate a set of realisations $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$. In rare cases we make use of bar notation to indicated a sequence (e.g. $\overline{\boldsymbol{X}}_K = \{X_1, \ldots, X_K\}$, though we point this out in context. We use enumerated subscripts to index items in a list. Where applicable, we use lettered subscripts to refer to a specific realisation. For example, if we have a cat and a mouse, we may refer to the cat's action as a_c, where a refers to the variable being an action realisation, while c denotes that it is the $(c)at's$ action selection.

Calligraphic letters (e.g. $\mathcal{N}$) are usually avoided in this text, but at times we use this to distinguish from bold letters. This is usually done for sets of sets, statistical distributions, or in general for clarity.

In general we use normal (round) brackets $(\cdot)$ in a variety of common ways, including to denote function arguments or tuples. Angled brackets $\langle\cdot\rangle$ are used, where otherwise not clear, to denote an ordered tuple or set. Curly brackets $\{\cdot\}$ are typically used for sequences or sets.

Table 1: Key Notational Conventions

Notation	Tag	Explanation
$\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$	Set of Realisations	Bold symbols denote a set of realisations.
Enumerated subscripts	Indexed Items	Index items in a list.
Lettered subscripts (e.g., a_c)	Specific Realisation	Refer to a specific realisation.
Calligraphic letters (e.g., $\mathcal{N}$)	Sets of Sets, Distributions, or Clarity	Distinguish from bold letters, denote sets of sets or statistical distributions.

Table 2: Key Notational Elements in Graphical Causal Inference

Notation	Tag	Explanation
V	Vertices (Nodes)	Variables in the system.
E	Edges	Dependence between variables.
$G = (V, E)$	Directed Acyclic Graph	Graph of vertices and directed edges, no cycles.
$\mathrm{Pa}(v_i)$	Parents	Directed edges to v_i.
$\mathrm{Ch}(v_i)$	Children	Incoming edges from v_i.
$\mathrm{An}(v_i)$	Ancestors	Vertices reaching v_i.
$\mathrm{De}(v_i)$	Descendants	Vertices reachable from v_i.
d-sep$(X; Y\|Z)$	d-Separation	Conditional independence criterion.
$do(X = x)$	Intervention	Setting X to x.
$Y_x(u)$ or Y_x	Counterfactual	Y under intervention $do(X = x)$.
$P(Y\|do(X = x))$	Causal Effect	Distribution of Y under intervention $do(X = x)$.
$G_{\overline{T}}$	Graph Modification	G where outgoing edges T are removed.

We also make use of standard notation and mathematical concepts applied in reinforcement learning. This will be useful in later chapters.

Table 3: Key Notational Elements in Reinforcement Learning

Notation	Tag	Explanation
s	State	A specific situation in the environment.
a	Action	A specific move or decision by the agent.
r	Reward	Scalar feedback signal received by the agent.
S	State Space	Set of all possible states.
A	Action Space	Set of all possible actions.
R	Reward Function	Function mapping state-action pairs to rewards.
T	Transition Function	Function mapping state-action pairs to state distributions.
γ	Discount Factor	Factor to discount future rewards.
π	Policy	Strategy followed by the agent.
$V^{\pi}(s)$	State-Value Function	Expected return from state s following policy π.
$Q^{\pi}(s, a)$	Action-Value Function	Expected return from state s, taking action a, following policy π.
ϵ	Exploration Rate	Rate of exploration vs exploitation.

Chapter 1

Introduction

"All reasonings concerning matter of fact seem to be founded on the relation of cause and effect. By means of that relation alone we can go beyond the evidence of our memory and senses." - David Hume [1].

Hume's insights into cause and effect resonate with the challenges in modern causal inference and reinforcement learning (RL). His assertion underscores a fundamental principle: mere correlation and data accumulation are insufficient for comprehending the complexities of the world. This thesis explores this principle in the context of machine learning (ML), where the maxim "data is not enough" highlights the limitations of current data-centric approaches in continuous learning algorithms.

The objective of this thesis is to bridge theoretical concepts from causality with practical methods and theory in RL. This endeavour begins with a comprehensive background on causal inference and causal learning, followed by an exploration of RL. Subsequent chapters synthesise these domains, delving into their intersection and the resulting methodologies.

An examination of contemporary trends and methods inspired by the intersection of RL and causality is also presented. This discussion aims to identify open questions in the field and propose methodologies for addressing them. Finally, the thesis culminates in a reflection on the initial research questions, discussing the outcomes and implications of this investigation.

1.1 Problem Area, Research Questions, and Hypotheses

The main focus of this research is how the study of RL and RL agents intersects with the study of causality. Special interest is taken to consider how methods and techniques from causal inference can assist in improving RL methods. The focus of this work is not on experiment, but rather on identifying trends and similarities in methods of research of seemingly similar fields. Where possible, we will make use of experiment and observation of performance to motivate arguments.

The key questions we investigate are as follows:

RQ1. Do existing reinforcement learning methods exhibit causal understanding?

RQ2. Does a causal model improve the sample efficiency and/or coordination of learning agents in a decentralised learning task?

RQ3. Can learning a causal model and applying it for applied causal inference lead to disparate impacts on sensitive subgroups?

The overarching hypothesis is that methods of modelling and reasoning in the fields of causal inference and RL have many intersections, albeit framed in different ways. We specifically hypothesise that methods from causality can aid in improving the reasoning ability of learning agents, especially with regards to improving sample efficiency and explicit counterfactual reasoning. Finally, we expect to find that this area is under-explored and warrants further investigation.

In terms of specific hypotheses pertaining to the identified research questions;

H1. We expect that there will be a range of causal abilities exhibited by different RL and MARL algorithms. That said, we hypothesise that classical algorithms will fail to exhibit the performance expected from a true counterfactual reasoning agent.

H2. We hypothesise that adding a causal model and/or causal methods to RL agents will improve the sample efficiency and rate of skill acquisition in RL environments.

H3. We hypothesise that applying a learnt causal model to decision tasks can lead to disparate impacts and biased outcomes.

1.2 Assumptions and Prerequisites

Though we attempt to be complete and avoid assuming too much preexisting knowledge of various fields, for brevity we make use of references and some assumptions of existing knowledge. We assume basic university mathematics, most specifically general knowledge of statistical distributions, the basics of calculus, as well as basic knowledge of computation and neural networks. For the most part we introduce required theorems and definitions, though we do not necessarily do this in the most introductory way. In this case, we refer the reader to popular sources and attempt to be consistent with these sources. In this way, this thesis should complement the selected resources and vice versa. Where something is not otherwise clear, please contact the author at uct@stjohngrimbly.com.

1.3 Thesis Structure

This thesis adopts a non-traditional structure, primarily in response to the broad scope of its subject matter, which bridges the gap between causality and RL. The broad scope has necessitated a comprehensive literature review and an extended introduction to vital concepts. Furthermore, in a departure from conventional academic layouts, the methodology and results sections have been merged. This merger ensures a coherent presentation, especially when interpreting results arising from closely related lines of inquiry.

Organised into seven chapters, the structure transitions from philosophical foundations to empirical investigations and theoretical explorations in the domains of causal inference and RL. It is formatted so as to provide a coherent narrative, guiding the reader through the evolution of key ideas and findings. The relationships between the chapters and the centrality of causality in this thesis are further illustrated in Figure 1.1.

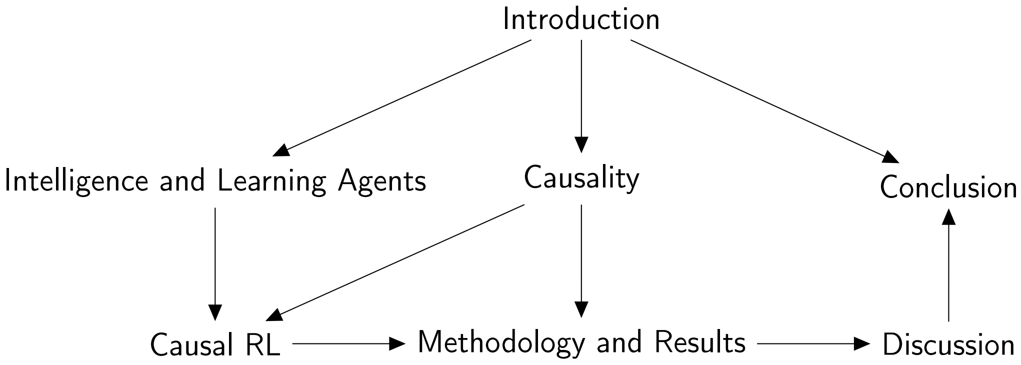


Figure 1.1: Causal graph illustrating the structural relationships among thesis chapters. While all chapters are interrelated, strong dependencies are denoted by causal edges, highlighting the central role of causality in this thesis.

1. **Introduction**: This chapter introduces the importance of understanding cause-effect relationships in today's data-centric world. It outlines the main research questions and provides an overview of the following chapters.

2. **Causality**: Discusses key concepts in causal inference and learning, essential for the research questions. Focuses on graphical methods and their connection to active learning.

3. **Intelligence and Learning Agents**: Explores the development of machine intelligence from Turing's predictions to modern Reinforcement Learning (RL). Examines how intelligence, RL, and causality intersect, setting the stage for later chapters.

4. **Causal Reinforcement Learning (CRL)**: Examines the combination of causal inference and RL, highlighting recent research in this field. Reviews methodologies and significant scholarly work in CRL.

5. **Methodology and Results**: Presents the research methods and detailed processes used. Showcases results from the sublines of investigation, including an investigation into how applied causal discovery can lead to unfair outcomes.

6. **Discussion**: Reflects on the methods and findings, with a focus on causal inference and (multi-agent) RL. Discusses integrating causal approaches in MARL, covering data fusion and counterfactual reasoning, and suggests future research areas. Discusses the results and implications of the investigation into bias and fairness of applied causal discovery in ML.

7. **Conclusion**: Summarises answers to the research questions, noting the potential usecases of causal models in RL and MARL. Highlights potential improvements in reasoning, efficiency, and fairness in learning algorithms from combining causal

inference and RL. Highlights the important issue of induced unfairness and disparaties on sensitive subgroups that can arise when incorrectly and naively applying causal discovery methods for applied ML projects. Suggests future research paths, including experimental validations.

1.4 Prior Works Incorporated in this Thesis

The foundation of this thesis is laid upon a collection of rigorous studies and projects I have either led or been a significant part of during my academic journey. These works not only enriched my understanding and insights into the domain but also serve as substantial contributors to the arguments, methodologies, and conclusions of this thesis.

1. **Causal Multi-Agent Reinforcement Learning: A Review and Open Problems**
   Published in the NeurIPS Cooperative AI Workshop in December 2021 [2], this paper provides an in-depth review of causal multi-agent RL. We explored the existing challenges and open questions in the domain. The paper's findings and methodologies serve as a cornerstone for several arguments in this thesis.

2. **Climbing the Ladder: A Survey of Counterfactual Methods in Decision Making Processes**
   My honours thesis [3], this comprehensive survey dives deep into the counterfactual methods employed in decision-making processes. The insights drawn from this survey not only shaped my perspective on the topic but also influenced the methodologies and arguments adopted in this current thesis.

3. **World Models and Predictive Methods in Deep Reinforcement Learning: A Survey**
   My honours RL project [4], this paper explores advanced model-based reinforcement learning techniques, integrating ideas from neuroscience and discussing deep learning and RL problems. It focuses on deep model-free and model-based methods, theories of latent representation, and predictive methods, culminating in an analysis of state-of-the-art models like MuZero.

4. **Mava: A Research Framework for Distributed Multi-Agent Reinforcement Learning**
   During my tenure at InstaDeep in 2021, I had the privilege to contribute to this project and paper. Serving as a whitepaper for the original Mava framework [5], it serves to introduces a research framework tailored for distributed multi-agent RL. Elements from this work have been integrated into various sections of this thesis to enhance its depth and relevance.

5. **Causal Bias and Fairness**
   My recent project undertaken during my internship under the supervision of Prof. Ferdinando Fioretto at Syracuse University in 2023, this ongoing study is looking into the intersections of causality, bias, and fairness. While the project is still in its development phase, preliminary findings and insights have been incorporated into this thesis, providing a novel perspective on the topics at hand.

Through this thesis, I strive to cohesively intertwine the learnings from these significant works, presenting a well-rounded perspective on *trends, problems, and solutions in causality and reinforcement learning.*

Chapter 2

Causality

This chapter develops and expands on the theory identified as necessary or important to investigate and answer the key questions identified in Chapter 1. Specifically, key concepts in causal inference and causal learning are introduced, with a specific focus on graphical methods and related research. This theory is derived and adapted from various sources, though large inspiration – as in the field itself – is drawn from Pearl [6]. This is not meant to serve as a complete or self-standing introduction to the wide fields of causal inference and causal learning, but rather it should suffice as both an introduction and a central hub of useful resources. The focus is placed on establishing ideas that relate causality to methods we will discuss in Chapters 3 and 4.

2.1 Introduction

Though almost a platitude at this point, the phrase "correlation does not imply causation" is one that bears repeating. Humans have an innate understanding of what cause-effect relationships are [7]. The problem, however, is that determining these relationships can be extremely challenging, especially when there is insufficient or conflicting information available from which an intelligent agent can draw conclusions. This shows up, for example, in (seemingly) irrational human behaviour [8].

The challenge of addressing counterfactual queries has been a persistent issue in the scientific discourse for centuries. A fundamental aspect, as identified by Karl Popper, is the principle of falsifiability, which is central to the validation of any scientific theory [9]. However, it is important to note that the reliance on falsifiability as a demarcation criterion for science is not without its criticisms, as highlighted by the Duhem–Quine thesis [10]. Attempting to attain certainty by contemplating what *could* have been presents a conundrum, given that the arrow of time progresses in a single, causal direction. Does this imply the exclusion of counterfactuals from scientific inquiry? This perspective was indeed shared by many eminent statisticians, including Fisher, particularly evident in debates such as "does smoking cause cancer?" [11–13]. There is speculation that an unobserved genetic factor might contribute to both a propensity for smoking and the development of cancer. The impracticality of conducting a randomised control trial (RCT) stems from various factors, including ethical, logistical, and practical challenges [14], necessitating the pursuit of alternative analytical methods. Are we, then, left incapacitated in reaching definitive conclusions due to the inherent limitations of our statistical approaches?

In addition to this, people often assume that cause and effect is an “all-or-nothing” endeavour. This is not strictly true, of course, since humans often assign cause-and-effect relationships to complex and uncertain scenarios. For example, it would not be uncommon to claim that “reckless driving causes accidents”[15]. By this we mean that reckless driving would be an *important* causal factor, if an accident were to occur. This is not to say that any time one drives recklessly, they *will* crash. This motivates the need for a theory which can handle this uncertainty in reasoning about causality, especially in terms of a causal hierarchy.

The concept of causation as articulated by Pearl is divided into three distinct levels in his *ladder of causation* model [16, 17]. At the base is the *seeing* level, which is focused on statistical correlations. The next level, *doing*, encompasses interventions and includes advanced methodologies like RCTs and RL techniques [18, 19]. In the RL context, an intervention is akin to an agent’s *action* that modifies the natural course of events, leading to outcomes such as new states and rewards.

We provide formal definitions of interventions and counterfactuals below. The mathematical details will become clear in this chapter.

Definition 2.1.1 (Intervention)**.** *Within a Structural Causal Model (SCM) M, an intervention I is the replacement of a set of structural functions in M with a new set. When intervention targets variable X_k, it is redefined as $X_k = \tilde{f}(\tilde{PA}_k, \tilde{U}_k)$, where $\tilde{PA}_k$ represents the new parental variables in the updated Directed Acyclic Graph (DAG). This alteration in causal mechanisms results in a novel interventional distribution, $P_{do(I)}$, and its corresponding probability density, $p_{do(I)}$.*

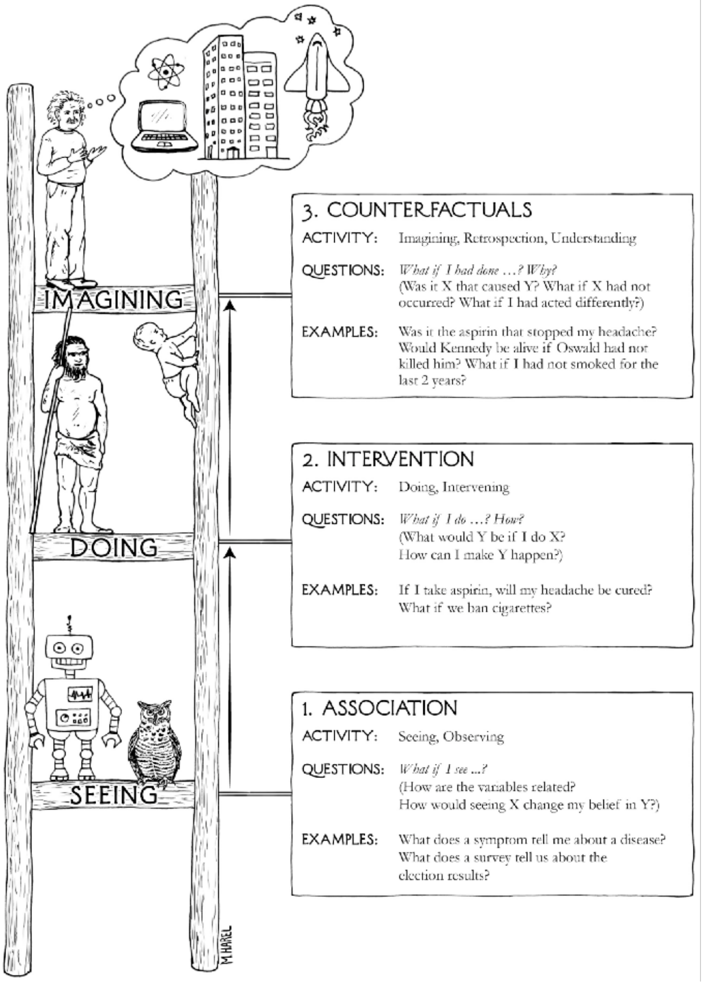
Surpassing the intervention level is the *imagination* stage, focused on counterfactual reasoning. Counterfactuals are hypothetical scenarios that have not actually taken place, and are fundamental to causal inference.

Definition 2.1.2 (Counterfactual)**.** *In an SCM $M = (S, P_{U_j})$ that encompasses nodes X with observed values x, a counterfactual SCM is formulated by substituting the noise variable distributions with those corresponding to the actual observed values $\boldsymbol{X} = \boldsymbol{x}$. The new noise distribution is denoted as the conditional probability $P_{U|X=x}$.*

In RL, counterfactual thinking is inherent, as agents might consider the outcomes of alternate actions not taken. This approach allows for a detailed examination of every possible action an agent *does*, *can*, or *could* have chosen. The intricacies of this conceptual ‘ladder’ are further developed in Pearl’s Causal Hierarchy [19], clarifying the formal mathematical relationships among these different types of data interactions.

2.2 The Fundamental Problem

Consider an individual, Casper, who is part of a new drug trial testing mRNA HIV vaccines. The important causal relationship is the effect of the treatment T on Casper’s immunity $Y_i(T)$. This notation denotes Casper as the individual i in the study, having an immunity measure Y after treatment T. We can represent the *potential outcome* by $Y_i(t)$. This is a *potential* outcome as the outcome is not considered with respect to a



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

Figure 2.1: Derived from [16], this figure illustrates Pearl's *Ladder of Causality* meta-model. It delineates the three tiers of causal reasoning: seeing, doing, and imagining, which relate to association, intervention, and counterfactual processes, respectively. The diagram provides examples for each level, aiding in the intuitive understanding which is helpful throughout this thesis.

specific individual. Assuming Casper receives the treatment $T = 1$, the experimenters are interested in the causal effect as measured by how much more immune Casper is *after* taking the treatment than if he *had not* taken the treatment, $T = 0$. We write this query as a difference in the *potential outcomes*, known as the *individual treatment effect* (ITE), denoted $\tau_i = Y_i(1) - Y_i(0)$. *The Fundamental Problem of Causal Inference* [20] is highlighted in this example - you cannot observe every potential outcome. What has happened, has happened.

The impossibility of determining ITEs in general is often dealt with by considering population level effects. Statisticians deal with this by considering distributions and using key statistics, such as the mean, to make predictions. In the same vein, the *average treatment effect* (ATE, or average causal effect (ACE)) is the expected difference in the potential outcomes with respect to individuals i, $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$. By linearity of expectation

$$\mathbb{E}\left[Y_i(1) - Y_i(0)\right] = \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] \overset{?}{=} \mathbb{E}\left[Y \mid T = 1\right] - \mathbb{E}\left[Y \mid T = 0\right],$$

where we have used $\overset{?}{=}$ notation to indicate careful consideration of this mathematical step. Unfortunately, things are not so simple with causal quantities (see Appendix A.2.4). In this case, we have attempted to replace causal quantities with associational quantities without considering what this means in the real world. Consider Figure 2.2 where X confounds the causal relationship between T and Y. Clearly, the ATE cannot be considered unless one considers the influence of X on the other variables. Conditioned on T, the influence of X on T is removed, but X still has an effect on Y.

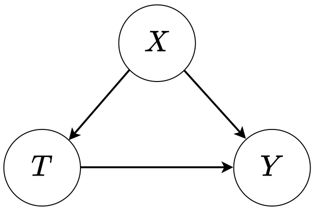


Figure 2.2: Example of a confounded relationship between variables in a causal structure. Here, X has a direct causal influence on both T and Y. Additionally, T has causal influence on Y.

2.3 Terminology & Formulation

This thesis primarily focuses on the Bayesian formulation of probabilities and conditional probabilities, in line with standard practices in graphical causal inference literature. This literature often examines variables in relation to each other. In this context, it is usual to condition variables on related variables, though sometimes these conditional variables may be left out for simplicity if it does not cause confusion.

Consider the variable A, symbolising the statement "St John will complete his MSc thesis in 2023." In this case, $P(A \mid K)$ indicates the subjective probability of A being true, based on a set of knowledge K. This notation and approach are central to the analysis and understanding of the causal relationships investigated in this thesis.

In keeping with statistical norms, random variables in this thesis are represented by capital letters. For example, V denotes a finite set of discrete random variables essential to our causal analysis. Each variable $X \in V$ is part of a domain D_X. Variables are denoted in uppercase (e.g., X, Y, Z), and their realisations in lowercase (e.g., x, y, z). Therefore, the probability of a variable taking a specific value is expressed as $P(X = x) = P(x)$.

As conditional independence is such an important idea in graphical (causal) modelling, we provide a definition here.

Definition 2.3.1 (Conditional Independence)**.** *Let $V = \{V_1, V_2, \dots\}$ represent a finite set of variables. Considering the joint probability distribution $P(V)$, for any $X, Y, Z \in V$, X is conditionally independent of Y given Z if $P(x \mid y, z) = P(x \mid z)$ whenever $P(y, z) > 0$. This means that knowing Z, Y does not add any extra information about X. This principle, vital to our study of causal relationships, is commonly expressed as $X \perp\!\!\!\perp Y \mid Z$.*

This section aims to provide a clear foundation for the probabilistic and causal frameworks underpinning the analysis in this thesis.

2.4 Graphs and Graphoids

In probability theory, considering conditional dependencies is fundamental for analysis. For example, when constructing a joint distribution over variables x_i, one may model this distribution as

$$\begin{aligned} P\left(x_1, x_2, \ldots, x_n\right) &= P\left(x_1\right) \prod_{i=2}^{n} P\left(x_i \mid x_{i-1}, \ldots, x_1\right) \\ &= P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_2, x_1)P(x_4 \mid x_3, x_2, x_1)\ldots \end{aligned}$$

This approach reflects a fundamental method for decomposing a multivariate distribution into a series of conditional distributions. The phrase "one may model" is employed here to indicate this methodological choice, highlighting a strategy for capturing the relationships among variables within a probabilistic framework.

By adopting this modeling technique, the complexity of modeling a distribution that encompasses an increasing number of variables becomes apparent as the requirement for parameters—based on which one conditions—grows exponentially. To elucidate, modeling $P(x_4 \mid x_3, x_2, x_1)$ necessitates an understanding of all 2^{n-1} possible permutations of the variables $\{x_3, x_2, x_1\}$. However, this complexity can be significantly reduced by taking into account the structure of the problem and focusing on only the local dependencies, thereby simplifying the modeling process as shown in Example 1.

Before extending these ideas to the *causal* domain, it is useful to consider the mathematical and structural similarities in statistical conditional independence and inference in undirected graphs. Pearl and Paz [21] do exactly this by showing that a shared set of axioms underpin both conditional independence and inference in undirected graphs. This theory forms a basis for extension to graphical-based inference and, by extension, graphical causal inference. Therefore, it is useful to briefly consider the defining aspects of this theory before discussing (graphical) causal inference. We refer the reader to Appendix A.1 for an in depth look at *graphoids*.

One can then define a mathematically rigorous version of the DAG, which will form the basis of (graphical) causal inference.

Definition 2.4.1 (Directed Acyclic Graph)**.** *A directed graph with no cycles is called acyclic and forms a* directed acyclic graph *(DAG).*

2.5 Graphical Models

Pearl [15, pg. 13] describes the role of graphs in statistical modelling as:

1. **Convenience**: Express substantive assumptions.
2. **Representation**: Economical representation of joint probability functions.
3. **Efficiency**: Facilitate inference from observations.

In terms of representation, graphical dependency models make decomposition of joint functions trivial. As we discussed earlier, we know that joint distributions can be written in terms of conditional dependencies. With the assumption that not every variable is dependent on each other, we can drop variables that are not important for modelling x_i. For example, say x_i is only sensitive to changes in some subset of the other variables, PA_i. Then we have that

$$P\left(x_i \mid x_{i-1}, \ldots, x_1\right) = P(x_i \mid pa_i).$$

Definition 2.5.1 (Markovian Parents)**.** *Let $V = \{X_1, \ldots, X_n\}$ be an ordered set of variables. A subset of variables PA_j is said to be the (Markovian) parents of X_j given that PA_j is the minimal set satisfying*

$$P(x_j \mid pa_j) = P(x_j \mid x_1, \ldots, x_{j-1}).$$

In other words, PA_j is sufficient to render X_j independent of all other predecessor variables.

The *local Markov assumption* says that given its parents in a DAG, a node X is independent of all its non-descendants. By non-descendants we mean all other nodes in the graph that are not in any directed path starting from X. This assumption makes the visualisation of independence a very powerful tool. The local Markov assumption is equivalent to Bayesian network factorisation, which implies that a joint distribution can be factored into a product of conditional distributions where all relevant information is taken into account. The minimality assumption adds that adjacent nodes in the DAG are dependent, which allows dependencies to be read off the directed graph.

The *Markovian parents* definition (Def. 2.5.1) relates to the fact that we can represent, in the form of a DAG, the sufficient dependency relationships between variables. We model the variables as nodes, and the edges are then representative of the dependencies between the variables. In fact, we can recursively construct the dependency DAG using the Markovian parents idea [15, pg. 15]. We start with a pair of variables, say (X_1, X_2), and draw an edge if they are dependent. We then proceed by finding a minimal set of Markovian parents, and draw an edge from each parent to the child node. In general, this results (recursively) in a DAG, called a Bayesian network (BN). That is, in a Bayesian network, an arrow from X_i to X_j assigns X_j via some functional dependency represented by the edge between them. By construction, a Bayesian network implies conditional independency via paths or lack of paths, a very useful and compact way to represent this information, especially as the scale of the DAG grows.

In the context of using DAGs, there are three fundamental node triplets to consider. These are chains, forks, and colliders.

In a chain $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$, the middle variable B can serve as a confounder, potentially opening a back-door path between A and C. Conditioning on or adjusting for B would close this path. In contrast, a collider $A \rightarrow B \leftarrow C$ poses a different challenge. If you condition on the collider B, it will actually open a back-door path between A and C. Therefore, to block all back-door paths, one must condition on the set of variables that includes confounders and ancestors of confounders but excludes colliders and descendants of the treatment variable. This provides a systematic way to identify the minimal set of variables required for back-door adjustment.

Example 1 (Simplified Likelihood)**.** *Imagine we work for university admissions and we are tasked with ensuring that our admissions process is fair. As domain experts we*

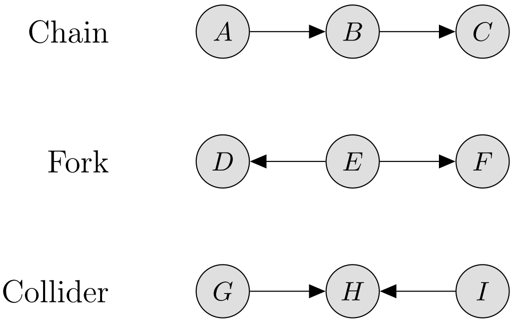


Figure 2.3: Graphical representations of three causal structures: (Top) A Chain, symbolising a sequence of events where A leads to B which then leads to C, much like links in a chain; (Middle) A Fork, where E acts as a source that splits its influence to both D and F, resembling the prongs of a fork; and (Bottom) A Collider, where G and I both direct their influence towards H, akin to two objects colliding into a common point.

have established five important measured factors - race R, income I, education quality E, SAT scores S, and college admissions C - that are predictors for college success.

As Bayesian thinkers, we decide to model these factors and their dependence relations as a graph. This allows us to form a DAG G, which is shown in Figure 2.4. We can simplify the likelihood over the graphical variables by considering the product of conditional distributions associated with each node in the DAG. Specifically,

$$L(\theta \mid x) = \prod_{\theta_i \in nodes(G)} P(\theta_i \mid pa_G(\theta_i)), \quad \forall \theta_i \in \{R, I, E, S, C\}$$

where θ_i represents the variables associated with node i, and $pa_G(\theta_i)$ denotes the set of parent nodes of θ_i in the DAG G. In this case, the variables are clearly dependent, as represented by the graphical structure. For the sake of this example, assume income is distributed normally with parameters μ and σ, while the other four factors are Bernoulli distributed with parameters p_j. The joint distribution on graph G is

$$P(R, I, E, S, C) = P(R) \cdot P(I \mid R) \cdot P(E \mid I, R) \cdot P(S \mid E) \cdot P(C \mid S).$$

Given a dataset with n students $\{x_1, \ldots, x_n\}$, the likelihood function is

$$\mathcal{L}(G) = \prod_{k=1}^{n} \prod_{j=1}^{5} P(x_{k;j} \mid \pi_{x_{k;j}}; \theta_j) \tag{2.1}$$

$$= \prod_{k=1}^{n} P(r_k; p_R) \cdot P(i_k \mid r_k; \mu_I, \sigma_I) \cdot P(e_k \mid i_k, r_k; p_E) \cdot P(s_k \mid e_k; p_S) \cdot P(c_k \mid s_k; p_C) \tag{2.2}$$

This likelihood formulation allows one to easily take the log-likelihood and observe that the full likelihood can be decomposed *into a sum of the conditional marginals.*

$$\ln \mathcal{L}(G) = \sum_{k=1}^{n} [\ln P(r_k; p_R) + \ln P(i_k \mid r_k; \mu_I, \sigma_I) + \ln P(e_k \mid i_k, r_k; p_E) + \ln P(s_k \mid e_k; p_S) + \ln P(c_k \mid s_k; p_C)]$$

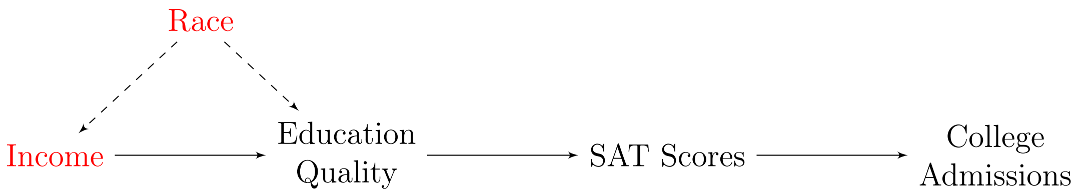


Figure 2.4: Causal diagram showing relationships among Race, Income, Education Quality, SAT Scores, and College Admissions. Nodes in red represent sensitive variables. Race is an unobserved confounder and serves as a latent sensitive variable. While not considered in this context, there could potentially be a direct, unfair edge from Race to College Admissions.

This simple example highlights some powerful outcomes that results as a function of assuming something about the structure of the problem. By using our domain knowledge to write down dependence relations, we were able to greatly simplify a full joint distribution into a product of marginal distributions. Keep this in mind for later, especially in regards to *causal discovery*.

Definition 2.5.2 (Markov Compatibility). *A probability function $P(\cdot)$ is considered Markov compatible with a DAG G if it can be decomposed such that $P\left(x_1, x_2, \ldots, x_n\right)=\prod_i P\left(x_i \mid pa_i\right)$, where each $P\left(x_i \mid pa_i\right)$ is in accordance with G. In this scenario, we state that G represents P, or alternatively, that P is Markovian with respect to G.*

The distributions compatible with a DAG G can be read off the DAG via the *d-separation* criterion. This is a foundational idea in the causal inference literature, much like that of independence in statistics. This, and related ideas, will be mentioned throughout this work.

Definition 2.5.3 (d-separation). *A set Z of nodes within a causal graph is considered to d-separate a path p under the following conditions:*

1. *If p includes any edge that passes through a vertex in Z, whether incoming or outgoing, or*
2. *If p involves a collision vertex not included in Z and this vertex does not have any descendants in Z.*

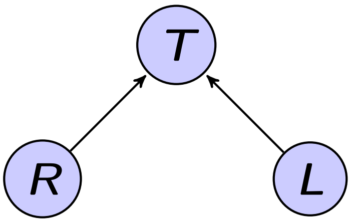
Sets X and Y in a causal graph are deemed d-separated by Z if all paths from X to Y are blocked by Z.

The idea behind d-separation is straightforward: we are trying to find the variables that, if conditioned on, would make two other (sets of) variables independent of each other. In other words, one is asking, "How would I block the dependence between these two variables?" Or rather, "How would I break this path of dependence?" This is not always as simple as it might first seem. For instance, in the causal diagram shown in Figure 5.4, consider the relationship between *Race* and *College Admissions*. According to d-separation, if we condition on variables like *Education Quality*, and *SAT Scores*, the direct path of dependence from *Race* to *College Admissions* could be blocked, implying these variables are d-separated in this context. However, an interesting caveat to consider is that of *Berkson's paradox* (see Example 20), where incorrectly conditioning on a variable can actually *introduce* spurious dependence between variables.

This highlights the nuanced nature of causal inference and the importance of careful consideration of the variables involved.

Example 2 (Rain, Traffic, and Leaving Late)**.** *Consider a DAG with three nodes: Rain (R), Traffic (T), and You Leave Late for Work (L). The relationships are defined as:*

- $R \to T$*: Rain can cause traffic.*
- $L \to T$*: Leaving late can lead to traffic.*



Using the d-separation criterion:

1. ***Without Conditioning**: R and T are dependent because rain can cause traffic.*
2. ***Conditioning on Traffic (T)**: If traffic is observed, R and L become dependent. This is due to the fact that conditioning on a collider like T can induce a relationship between its causes.*
3. ***Conditioning on Rain (R)**: Knowing it's raining, the path between L and T is blocked. L and T become conditionally independent given R.*

The key takeaway is that d-separation allows us to determine conditional independence relationships in a causal graph, providing a bridge between graphical causal models and statistical dependencies.

We discuss some more rigorous theory of DAG-induced graphoids and implications of conditional independence. Though this is not core to the argument, the reader is encouraged to read through Appendix A.1.1.

2.6 Causal Networks

One can augment graphical models with causal assumptions such that they become *causal* graphical models. Perhaps already an implicit assumption to the reader, a variable X is a *cause* of a variable Y if Y can be affected by changes in X. One element that makes reading causal structure off a directed graph convenient is the *strict causal edges assumption*, where every parent variable is assumed to be a direct cause of all child variables. It is important to note that this is not a necessary assumption for Bayesian networks to make sense. This assumed extension then forms a *causal Bayesian network* (Definition 2.6.1).

Pearl [15, pg. 21] makes a strong intuitive argument for why one would want to make these causal assumptions when they first appear not to be necessary for Bayesian inference. Simply put, when there exists dependencies between variables that aren't considered "causal," we think of them as spurious. To quote Pearl directly: "It seems that if conditional independence judgements are by-products of stored causal relationships, then tapping and representing those relationships directly would be a more natural and more reliable way of expressing what we know or believe about the world. This is indeed the philosophy behind causal Bayesian networks."

Perhaps more interesting at a technical level, another advantage of applying causal assumptions to Bayesian networks is that modifying the network when faced with new information is much more tractable. A local change in mechanism can be modified into 'an isomorphic reconfiguration of the network topology.' One can imagine how crucial this is to machine *learning* problems, especially where adaptation and *learning* is important.

2.6.1 Causal Networks as Oracles

Having changes in causal mechanisms only affect the local network topology is what allows the flexibility of causal inference and causal models. The modular nature of such a model allows one to reason about how local changes would change outcomes without *rethinking* or *retraining* the model as a whole. Once again, this is a fundamental point. A causal model (e.g. CBN) is much more informative than a statistical model (e.g. BN without causal assumptions) because it allows us to reason about how probabilities would change due to external intervention. Graphically, an intervention corresponds to "deleting" an edge. This makes intuitive sense as an intervention removes the causal dependencies of a child node on its parents. Interventions of this nature are not easily represented in statistical language. Instead, we can denote an intervention via the $do(\cdot)$ notation, where the variable being intervened on is acted upon by the *do*-operator. This notation will be used thoroughly throughout this work.

Definition 2.6.1 (Causal Bayesian Network)**.** *Consider a probability distribution $P(v)$ across a variable set V. The distribution resulting from the intervention $do(X = x)$, which sets a subset of V to a fixed value x, is noted as $P_x(v)$. Let $P_\star$ be the collection of interventional distributions where no intervention is made, represented by $do(X = \emptyset)$. A DAG G aligns as a causal Bayesian network (CBN) with $P_\star$ if, for every distribution P_x within $P_\star$, the following conditions hold:*

(i) *The distribution $P_x(v)$ adheres to the Markov property with respect to G.*

(ii) *For each V_i in X, $P_x(v_i) = 1$ provided that v_i aligns with the condition $X = x$.*

(iii) *For all V_i not in X, $P_x(v_i \mid pa_i)$ equals $P(v_i \mid pa_i)$, given that pa_i is in agreement with $X = x$.*

This definition allows us to talk about all equivalent interventional distributions in an efficient manner while maintaining the factorisation properties of Bayesian networks. These conditions also imply that intervention and conditioning on parent variables are locally equivalent, $P(v_i \mid pa_i) = P_{pa_i}(v_i)$ - a simple yet powerful property. Further, variables are locally invariant to interventions s elsewhere in the model, $P_{pa_i,s}(v_i) = P_{pa_i}(v_i)$. This fundamentally distinguishes causal models from statistical models.

2.7 Functional & Structural Causal Models

Functional deterministic models, foundational in causal modeling's early history, remain influential in social and economic sciences. These models, exemplified by Wright's genetic inheritance studies [22], introduce uncertainty via unobserved (latent) variables. The observed randomness, in this view, stems entirely from unmeasured conditions. Such quasi-deterministic frameworks facilitate counterfactual reasoning, an aspect absent in stochastic models, to be discussed in depth later.

In a functional causal model, the equations $x_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n$, are key, with pa_i representing parent variables and U_i signifying noise/error from unknown factors. These equations, much like deterministic laws in physics, assign values to variables. An independent and autonomous mechanism for each equation constitutes a structural model. When every variable in the equations has a corresponding assigning function (appearing on the left-hand side), it forms a structural *causal* model.

A SCM uniquely facilitates explicit exploration of interventional and counterfactual questions. A counterfactual is depicted as an observed distribution for a certain covariate condition, say $X = x$, but conditioned on an alternate reality, $X = x'$. Using potential outcomes, Y_x represents the outcome under $X = x$. The counterfactual distribution is then $P(Y_x \mid x')$. Although a detailed discussion follows later, the key idea is that the observed outcome distribution under $X = x$ informs us about the model's 'noise'. This knowledge helps model outcomes under different conditions. Importantly, a causal model's falsifiability means it can be refuted if experimental data contradicts the model, confining its logic to the truthfulness of the underlying model.

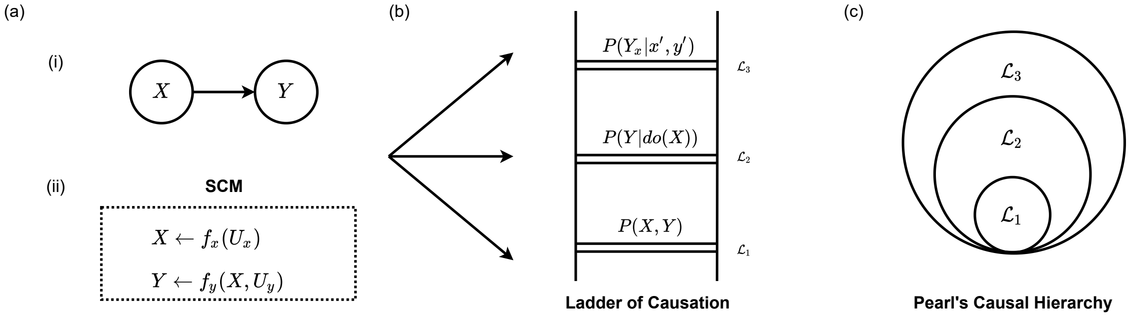


Figure 2.5: (a)(i) A *causal Bayesian network* displayed as a DAG, showing causal relationships between variables. **(ii)** A SCM aligns with the DAG from (i), having two main variables: exogenous X, influenced by external noise U_x, and endogenous Y, affected by X and its noise factor U_y. **(b)** Different observable phenomena originating from an SCM. The SCM generates three causal quantities, each corresponding to a level on the 'ladder of causation'. Associational metrics are on $\mathcal{L}_1$, interventions (denoted as $do(X = x)$) on $\mathcal{L}_2$, and counterfactual metrics (envisioning alternate scenarios Y_x under $X = x'$) on $\mathcal{L}_3$. **(c)** A Venn diagram illustrating Pearl's Causal Hierarchy [19], highlighting the hierarchical arrangement from (b).

Definition 2.7.1 (Structural Causal Model [19]). *An SCM, represented as M, is defined by a quartet $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$, which includes:*

1. $\boldsymbol{U}$*: Exogenous variables, determined by external factors.*
2. $\boldsymbol{V}$*: Endogenous variables, affected by elements within the model.*

3. ***F***: *Functions linking* ***U*** *and* ***V****, where each* f_i *creates a connection from* $U_i \cup Pa_i$ *to* V_i *as* $v_i \leftarrow f_i(pa_i, u_i)$*. Here,* pa_i *belongs to* Pa_i*,* U_i *is part of* ***U****, and* Pa_i *is a subset of* $\boldsymbol{V} \setminus V_i$*, assigning values to variables based on their interrelationships.*

4. $P(\boldsymbol{U})$*: A probability distribution covering* ***U****.*

Example 3 (Structural Causal Model)**.** *Let's consider a simple system consisting of two variables: altitude* A *and temperature* T*. Our hypothesis is that altitude causally influences temperature, which means the temperature of a location is determined, in part, by its altitude.*

Causal Bayesian Network*:*

1. *We represent the causal relationship with a directed edge from* A *to* T*. This is based on the belief that an increase in altitude generally corresponds with a decrease in temperature.*

2. *Our SCM for this system can be represented as:*

   (a) $A = u_a$

   (b) $T = f(A, u_t)$

Here:

- A *is the altitude which is exogenous and determined by external noise factors, represented by* u_a*.*

- T *is the temperature which is endogenous, influenced by altitude* A *and some noise factor* u_t*.*

Representative Equations*:*

$$A = u_a$$
$$T = \beta \cdot A + u_t$$

Explanation*: Imagine two locations: one at sea level (altitude = 0) and another on a mountain top (altitude = 2000 meters). Our model might predict that the mountain top location is cooler than the sea level location because of the difference in their altitudes. It's important to emphasise that our belief is in the direction from altitude to temperature, and not vice-versa. We wouldn't expect an intervention that changes temperature to consequently change the altitude.*

SCM Visualisation*:*

Interventions*: If we were to perform an intervention where we artificially set the altitude (a hypothetical scenario, of course), the resulting temperature would change according to our model. This can be represented as* $do(A = a)$*, where we set the altitude to some value* a *and observe the resulting temperature. Notably, if we were to perform an intervention that artificially changed the temperature, we would not expect this to result in any change to altitude, consistent with our belief about the direction of causality.*

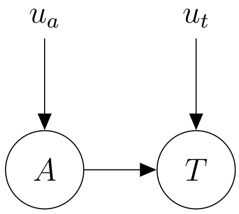


Figure 2.6: A SCM representing the causal relationship between altitude A and temperature T. Here, altitude is considered as an exogenous variable influenced by an external factor u_a, while temperature is an endogenous variable influenced by altitude and another external factor u_t.

2.8 Procedure of the Causal Inference Engine

Judea Pearl's concept of a *Causal Inference Engine* effectively provides a systematic approach to perform causal inference. This 'engine' represents a series of steps for deriving causal effects from observational data, based on the theoretical framework of SCMs. The process is depicted in Figure 2.7 and involves the following stages:

1. **Model Specification:** Identify and define the variables and their causal relationships within an SCM. This initial step leads to the creation of a *Causal Model*, visualised as a DAG in Figure 2.7.
2. **Query Formulation:** Develop causal queries to define the *Causal Estimand*. These queries may focus on either interventional or counterfactual effects based on the established model.
3. **Identification:** Apply do-calculus rules to verify if the causal estimands can be identified from the available data. If identifiable, derive the identification formula, which leads to a *Statistical Estimand*.
4. **Estimation:** Using the identification formula, employ statistical methods to calculate the causal effects from the data. This process results in an *Estimate*.
5. **Refutation / Validation:** Conduct robustness tests, sensitivity analyses, and if possible, compare with experimental data to validate or refute the estimated causal effects and underlying assumptions of the model.

The procedure begins with *Model Specification* of causal relationships, proceeds to *Query Formulation* for targeting specific causal effects, moves to *Identification* to evaluate the possibility of extracting these effects from the data, and then to *Estimation* of the causal effects. The final step, *Refutation / Validation*, is crucial for verifying the accuracy and reliability of the deduced causal relationships. This structured approach ensures a thorough method for analysing and confirming causal effect estimations within the framework of SCMs.

Conventional model selection in statistics primarily revolves around the task of finding the model that best fits the observed data, often without an explicit focus on causal relationships. The main target is usually prediction accuracy, goodness of fit, or variance explanation within the data. Common approaches include hypothesis testing, likelihood ratio tests, and scoring using Akaike Information Criterion (AIC) or Bayesian

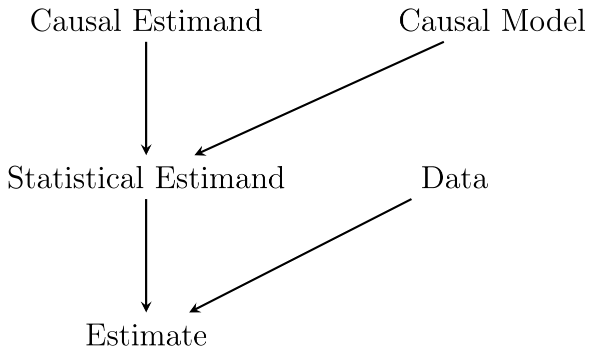


Figure 2.7: Diagram illustrating the process of causal estimation. The *estimand* represents what we wish to estimate — a formal definition of the causal quantity of interest. The *estimate* is the actual value approximating the estimand, derived from the data. The causal model provides a structured representation of our assumptions about the causal process, guiding the translation of the causal estimand into a statistical one.

Information Criterion (BIC), among others. These techniques generally do not differentiate between correlation and causation, and often operate under the assumption that the underlying data-generating process is stationary and devoid of interventions. Below, we contrast this conventional approach with the procedure outlined in the causal inference engine:

1. **Explicit Causal Focus:** Unlike conventional model selection, the causal inference engine begins with an explicit emphasis on causal relationships, grounding the analysis in a theoretical framework that allows for causal interrogations.

2. **Interventional and Counterfactual Queries:** The causal inference engine facilitates the examination of interventional and counterfactual scenarios, enriching the analysis beyond mere associative relationships often sought in conventional statistical modeling.

3. **Model Falsifiability:** As highlighted earlier, a distinctive feature of causal models is their falsifiability. If experimental data does not align with model predictions, the causal model can be refuted. In contrast, conventional statistical models often lack such clear falsifiability criteria.

4. **Robustness to Interventions:** Causal models, being designed with interventions in mind, are better suited for extrapolation under changing conditions or external interventions, a scenario where conventional statistical models might falter.

5. **Informative Model Specification:** The initial model specification in causal analysis is often guided by theory and domain knowledge, making it a more informed and structured approach compared to the data-driven nature of conventional model selection which might overlook underlying causal structures.

The Causal Inference Engine elevates the analytical rigour from mere pattern recognition or predictive modeling, common in conventional statistical model selection, to a level where causal relationships can be explicitly modelled, tested, and interpreted.

This transition augments the scope and depth of insights that can be derived, thus offering a deeper understanding of the underlying system dynamics and the implications of potential interventions.

At this point it is important to mention that the (Pearlian) graphical approach to causality is not the only approach. We include an introduction to some other frameworks, including the *potential outcomes* and *information theoretic approach* in Appendix A.2.

2.9 Intervention, Identifiability, and a Calculus of Interventions

At this point, we have established several core concepts in causal inference. We have defined SCMs and related necessary concepts, and shown how it can be used to generate different types of causal quantities. We have also discussed the idea of an intervention, and how it can be used to generate counterfactual quantities. However, we have not yet discussed how to use these models for estimation and inference. In this section we will discuss the idea of identifiability, and extend this to a complete *calculus of interventions.* This calculus will allow us to infer which causal quantities can be estimated from different forms of observational and interventional quantities available, given a SCM. For those not familiar, we introduce the idea of *causal quanities* in Appendix A.2.4.

2.9.1 Identification of Causal Quantities

Identifying these various causal quantities is non-trivial, primarily because they are defined in the context of a specific causal model, as opposed to a joint statistical distribution. The challenge arises in disambiguating multiple observational models that could potentially be generated by different underlying causal structures. The concept of *identifiability* serves to alleviate this issue, imposing constraints that allow us to uniquely estimate causal quantities, given a set of assumptions embedded in a causal model M. The identifiability of causal quantities is essential for meaningful causal inference. It ensures that given a specific causal model M and sufficient observational data, the causal quantities - whether it's ATE, ITE, ATT, TE, or CATE - can be reliably estimated, thus bringing our theoretical constructs into the experimental and verifiable domain.

Definition 2.9.1 (Identifiability)**.** *Consider a computable quantity Q of a model M within a specified class of models M. The quantity Q is deemed identifiable in the model class M if, for any two models M_1 and M_2 in M, the condition $Q(M_1) = Q(M_2)$ is met whenever the probabilities $P_{M_1}(v) = P_{M_2}(v)$ are equal. In cases where only a limited observation set is available, providing a subset of features F_M from $P_M(v)$, the identifiability of Q relative to F_M is established if $Q(M_1) = Q(M_2)$ is satisfied whenever $F_{M_1} = F_{M_2}$.*

Considering this in the context of learning distributions from data (e.g. in ML), one ideally would want a way to compute quantities from observational data. This sounds obvious as this is, perhaps implicitly, the goal of much of ML. The non-trivial component comes in when the goal shifts to making *causal* predictions rather than correlative (statistical) predictions.

Imagine there is a quantity Q from which one would like to determine the causal effect on y given knowledge $\hat{x}$ under model M, written $P_M(y \mid \hat{x})$. Given a fully specified model M, this is certainly computable. The problem is, models usually have some incomplete element - perhaps some noise on measured variables, or other latent variables in the model. This is exactly the case where the notion of identifiability becomes useful.

Initially, two fundamental assumptions about the problem's structure are necessary. The models in question must:

(i) Possess a common causal graph G;

(ii) Yield positive probability distributions for observed variables, denoted as $P(v) > 0$.

Definition 2.9.2 (Identifiability of Causal Effect)**.** *For a specified graph G, the causal impact of variable X on variable Y is deemed identifiable when the probability $P(y \mid \hat{x})$ can be uniquely determined based solely on the probabilities of observed variables. Formally, this means that for any pair of models M_1 and M_2 with a shared probability distribution $P_{M_1}(v) = P_{M_2}(v) > 0$ and the same graph $G(M_1) = G(M_2) = G$, the relationship $P_{M_1}(y \mid \hat{x}) = P_{M_2}(y \mid \hat{x})$ consistently holds.*

With these assumptions and the definition in place, identifiability signifies that deducing the causal effect of an action $do(X = x)$ on Y is feasible when provided with (1) observational data, and (2) a causal graph that delineates the variables in the data generation process.

The stipulation for positive distributions simplifies the issue by precluding scenarios where $X = x'$ has a zero probability, which would otherwise render inferring the effect of action $do(X = x')$ unfeasible. To clarify the concept of variable inclusion for identifiability, we present the following theorem:

Theorem 2.9.1. *For any Markovian model with a causal diagram G, where a subset V of variables is observed, the causal effect $P(y \mid do(x))$ is identifiable if the observational dataset includes the variables X, Y, and all parent variables of X (denoted as PA_X). In formal terms, identifiability is ensured if the observational dataset encompasses all variables in the set $\{X, Y\} \cup PA_X$, meaning that X, Y, and PA_X are all part of the observed variables set V.*

2.9.2 Adjustment Criteria and Formulae

Dealing with confounding is foundational to experimental design and, of course, the study of causality. The influence of confounding factors can make or break experiment and analysis. The when and how of adjusting for these factors is not always obvious. Simpson's paradox provides some clear examples of why dealing with confounding is essential for many statistical problems. To explain simply, Simpson's paradox occurs whenever inclusion of additional factors reverses a statistical relationship. At first glance this seems implausible - surely more information would only serve to increase the statistical 'certainty' one has about the model?

Simpson's Paradox Example Consider the following: you have a kidney stone and are consulting a doctor about treatment options. You are faced with two treatment options (1) treatment A and (2) treatment B, but you would like some data to make an informed decision as to which treatment is better for your circumstances. The doctor provides you with the following tabulated data:

Table 2.1: Comparison of Recovery Percentages for Two Treatments (A and B)

	Treatment A	Treatment B
Recovery %	0.78	0.83
Total Recovery %	0.80	

From this you decide, quite reasonably, that your best choice of treatment is treatment B. However, the problem is that there are some missing data. Consider the exact same dataset, but now with additional context about how these data were generated. In table 2.2 we see that there are two different classes of kidney stone sizes. Suddenly it appears as though treatment A is better than treatment B, since it is better over both small and large kidney stone strata. If this appears confusing, focus on the denominators and notice that structural difference between assignment of treatment groups.

Table 2.2: Comparison of Treatment Success Rates for Small and Large Kidney Stones with Treatments A and B. Example taken from Peters et al. [23].

	Treatment A	Treatment B
Small Kidney Stones	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Large Kidney Stones	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
Total Recovery	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$

The kidney stone example involves only three interacting variables, with a very simple graph structure. One can imagine how much more difficult it can become to deal with these paradoxes, especially as the problems scale in dimension. This motivates the need for tools that (1) help identify when and how confounding is an issue, and (2) a way to adjust for these confounding variables. This idea of 'adjusting' being important for establishing causal effects is what leads to various criteria - including the back-door and front-door criteria. We refer the reader to Appendix A.3, which serves as a foundation for what follows.

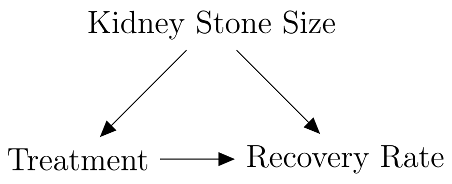


Figure 2.8: Causal DAG representing relationships between kidney stone size, treatment, and recovery rate. This structure can result in Simpson's paradox as observed in Table 2.2.

2.9.3 Calculus of Intervention

Pearl's rules of do-calculus provide a set of principles for manipulating causal expressions involving interventions, formalised within the framework of SCMs. Comprising three rules, the do-calculus allows for the translation of causal queries into expressions that can be evaluated from observed data, under certain identifiability conditions. Intuitively, the rules exist to simplify the causal analysis by identifying when certain variables or interventions can be ignored, marginalised, or treated as observed.

These rules are instrumental in making causal inferences more tractable and facilitating the estimation of causal effects from empirical observations, even when controlled experiments are not feasible. By reducing complex causal structures to simpler expressions, the do-calculus contributes significantly to the advancement of causal reasoning and analysis in various scientific disciplines. More formally Pearl's *do*-Calculus allows one to *identify* any causal quantity that is identifiable. This fact has been proven since it has been shown that *do*-Calculus is complete for identification. We now present the rules and then walk through what these equations mean in the context of the previous discussions.

Rule 1: $P(y \mid do(t), z, w) = P(y \mid do(t), w) \quad \text{if } Y \perp_{G_{\overline{T}}} Z \mid T, W$

Intuitively, this rule says that if there is no direct causal path from Z to Y (given T, W), then knowing Z doesn't change the distribution of Y given an intervention on t - $do(t)$ - and a known w. This is a simple extension of d-separation to interventional distributions.

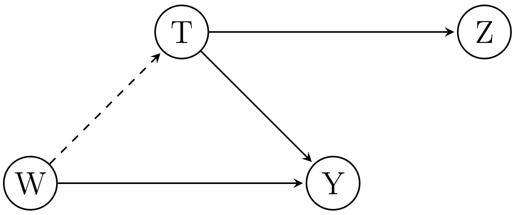


Figure 2.9: Causal DAG with Treatment (T), Outcome (Y), Confounder (W), and an unrelated variable (Z).

Rule 2: $P(y \mid do(t), do(z), w) = P(y \mid do(t), z, w) \quad \text{if } Y \perp_{G_{\overline{T},\underline{Z}}} Z \mid T, W$

This rule applies when there is a joint intervention on T and Z. Essentially, if the causal effect of Z on Y disappears when conditioning on T, W, then the intervention on Z becomes redundant. This is a generalisation of the backdoor criterion.

Rule 3: $P(y \mid do(t), do(z), w) = P(y \mid do(t), w) \quad \text{if } Y \perp_{G_{\overline{T},\overline{Z(W)}}} Z \mid T, W$

This rule also involves joint intervention on T and Z, but here Z depends on W. This rule implies that if the causal effect of Z on Y disappears when (1) conditioning on both T and W, and (2) Z depends on W, then the interventions on Z and the inclusion of W in the conditioning become redundant. That is, we don't want to incorrectly condition on node(s) Z that are ancestors of W.

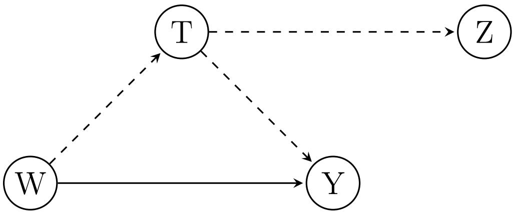


Figure 2.10: Causal DAG with Treatment (T), Outcome (Y), Confounder (W), and an intervened variable (Z, with a dashed line representing the intervention).

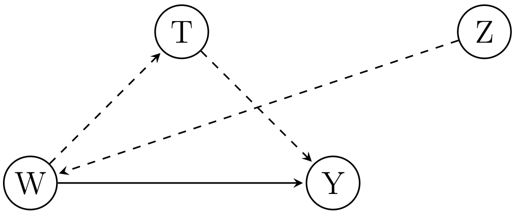


Figure 2.11: Causal DAG with Treatment (T), Outcome (Y), Confounder (W), and an additional variable (Z) that depends on W. The dotted edges from W to T and from W to Z represents a previous edge that was removed via intervention on the parent variables, and the graph satisfies the condition for the third rule of do-calculus.

2.10 Path-Specific Causal Effects

Understanding path-specific causal effects is essential for dissecting complex causal relationships. Unlike conventional causal analysis that considers total causal effects, path-specific effects isolate the contribution of specific causal paths [24–26]. By isolating specific causal paths, researchers can gain nuanced insights into the underlying mechanisms of causal systems [24]. This approach has proved to be useful in practice, finding significant importance in the study of causal fairness, especially in the context of healthcare [27].

Considering a causal structure represented by a DAG, let X be a treatment, Y an outcome, and M a mediator variable. We can represent the path-specific effect of X on Y through the mediator M as:

$$X \to Y_M = \mathbb{E}\left(Y \mid do(X=1), M\right) - \mathbb{E}\left(Y \mid do(X=0), M\right). \tag{2.3}$$

Example 4 (Effect of Education on Income)**.** *Consider a causal system where we are interested in the effect of education X on income Y, mediated by a variable such as job skill level M. The path-specific effect of education on income through job skill level can be quantified using the above equation.*

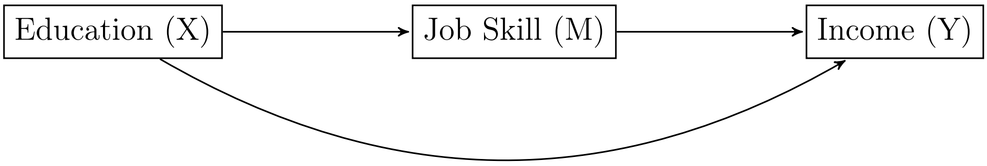


Figure 2.12: Causal structure of the effect of education on income through job skill level.

This figure represents the causal structure of our example, with two paths from X to Y: one direct and one through M. The path-specific effect isolates the effect of education on income through the job skill level [26].

As an aside, we briefly discuss the relationship between causal models and differential equation modelling in Appendix A.5, though this is not core to the argument of this thesis.

2.11 Causal Structure Learning

Though the theory of causal inference has come a long way in recent decades, causal structure learning has always been a harder problem to tackle. This is natural in that structure learning naturally involves going from effects to causes, or from correlations to causation, which is a fundamentally harder problem. This is effectively what is studied in much of ML, where the goal is to learn a predictive model of the world. The assumption that is implicitly made by ML practitioners is that applying ML methods to Big Data will reveal ground truths from correlations in the data. As in causal inference, the key step that causality offers is the explicit consideration of how to go about learning given limited (incomplete) data. Of course, if one had data for every possible outcome, one would not need a causal model for prediction within the domain since a simple lookup would do. That said, if intelligent behaviour is the goal, a certain level of generalisation and abstraction is key, and lookup tables do not solve this problem. We discuss this in more detail in Chapter 3 as this pertains to **RQ1** and **RQ2**.

The process of structure learning can itself be scored by probabilistically considering how well the current model describes the data, much in the way that a statistical model is learnt. This learning process becomes 'causal' when the constraint that each relationship must be causal is implemented. For example, one may score a graphical model's description of the data generating process by comparing observed data with the predicted results of the model. This leads to several criteria and score functions that can then be optimised for.

Having made such a fuss about not learning causation from correlations, why would one try to do exactly that by attempting to learn a causal model from associational data? Consider the following thought experiment: can you think of a situation where variables A and B are dependent, B and C are dependent, but A and C are *not* dependent? No matter the details of what one comes up with, the *structure* that results will look something like $A \to B \leftarrow C$. i.e. A and C are independent causes of B. Although simple, simple common structures in causal graphs can allow uncovering of causal structure via simple methods of induction. This process of causal structure *discovery* is now discussed.

First, one can describe what a *structure* is in terms of a graphical structure. Intuitively, a structure should inform the modeller about the relationships between variables in a system.

Definition 2.11.1 (Causal Structure)**.** *A causal structure of a set V of variables is a DAG in which every node corresponds to a unique element of V and every edge corresponds to a direct functional relationship between the adjacent nodes (and therefore* *the variables).*

At a deeper level, one can use the structure to more fully describe a model by specifying what the relationships between the system variables are. This defines a Causal Model.

Definition 2.11.2 (Causal Model)**.** *A causal model is a pair $M = \langle D, \Theta_D \rangle$, where D is a causal structure, and Θ_D are parameters that assign to each variable $X_i \in V$ a function $x_i = f_i(pa_i, u_i)$, as well as a probability measure $P(u_i)$ to each u_i. As is the usual notation, PA_i denote parents of X_i, and U_i represent random independent noise variables.*

2.11.1 Model Preference

The flexibility to include an arbitrary set of variables V in a causal model presents a challenge. Specifically, an unlimited array of variables allows for an endless variety of representations for the same distribution. Consider a hidden causal variable V_i that is the root of the underlying causal structure. One could endlessly extend the model by adding variables that act solely as child nodes to V_i, effectively modelling any idiosyncrasies or noise in the data. This flexibility is reminiscent of the principle of universal approximation in neural network (NN) theory, where NNs with sufficient complexity can approximate any continuous function (see Section 3.3). Hence, both NNs and causal models share the capability to represent a wide range of complexities.

Given this potential for infinite complexity, a mechanism for model selection becomes imperative. One compelling strategy prioritises *minimal* models, a preference underpinned by Occam's Razor. Simpler models, being more easily falsifiable, are often favoured. According to this rationale, when faced with equally explanatory models, one oriented towards empirical verification would opt for the simpler alternative.

Definition 2.11.3 (Inferred Causation)**.** *A variable X exerts a causal effect on another variable Y if there is a direct path leading from X to Y present in all minimally complex structures that are compatible with the observed data.*

Definition 2.11.4 (Latent Structures)**.** *A latent structure, denoted as $L = \langle D, O \rangle$, is characterised by D, a causal framework over the variable set V, and O, a subset of V representing the observed variables.*

Definition 2.11.5 (Structure Preference)**.** *A latent structure $L = \langle D, O \rangle$ is preferred to another structure $L' = \langle D', O \rangle$ (expressed as $L \preceq L'$) if and only if D' is capable of replicating the behaviour of D over the observed set O. In formal terms, this is true if for every parameter set Θ_D, there exists a corresponding set $\Theta'_{D'}$ such that the probability distributions $P_{[O]}(\langle D', \Theta'_{D'} \rangle)$ and $P_{[O]}(\langle D, \Theta_D \rangle)$ are identical. Two latent structures are said to be equivalent (denoted as $L' \equiv L$) if both $L \preceq L'$ and $L' \preceq L$ hold true.*

Not to get too entangled with definitions, all that structure preference defines is a formalisation of the idea that *less is preferable* when considering the number of necessary variables.

Definition 2.11.6 (Minimality)**.** *A latent structure is regarded as minimal within a specified class $\mathcal{L}$ of latent structures if it is not strictly less preferred than any other member of $\mathcal{L}$. In other words, for every $L' \in \mathcal{L}$, $L \equiv L'$ is true whenever $L' \preceq L$.*

In simple terms, the only preference that can be placed on a minimal latent structure is the trivial equivalence. In this sense the minimal latent structure is the 'smallest' possible.

Definition 2.11.7 (Consistency)**.** *A latent structure $L = \langle D, O\rangle$ is consistent with a distribution $\hat{P}$ over the observed variables O if the causal framework D supports a model that can produce $\hat{P}$. Specifically, this is valid if there is a parameter set Θ_D such that the probability distribution $P_{[O]}(\langle D, \Theta_D\rangle)$ equals $\hat{P}$.*

The idea of consistency here extends the intuitive understanding of what consistency means across a wide range of fields. That is, there is another way to 'describe' (parameterise) the object, in this case, the latent structure. Consider how important this is if one would like to 'learn' a structure using some parameterised method. For example, using NNs.

Definition 2.11.8 (Inferred Causation (Extended))**.** *For a given distribution $\hat{P}$, a variable C is said to causally influence another variable E if and only if there is a sequence of directed links from C to E in each minimal latent structure that is in agreement with $\hat{P}$.*

2.11.2 Stability/Faithfulness

The concept of stability or faithfulness has emerged as a prominent theme in several areas of causality research. The idea is straightforward: a causal distribution should entail the same independence relations regardless of the particular parameterisation of the distribution. Let's consider a model that includes both causal paths: $X \to Z$ and $X \to Y \to Z$. This structure can be likened to the one shown in Figure 2.12, although in that figure we used different variable labels.

In this scenario, we define the following linear causal relationships: the direct effect of X on Z is positive with a strength of 1; the influence of X on Y is positive with a strength of 0.5; and the influence of Y on Z is negative with a strength of -2. Therefore, the mediated effect of X on Z via Y is $0.5 \times -2 = -1$, which exactly cancels out the direct effect of X on Z. This results in a total effect of $1 + (-1) = 0$, misleadingly suggesting independence between X and Z, despite a clear causal pathway through Y.

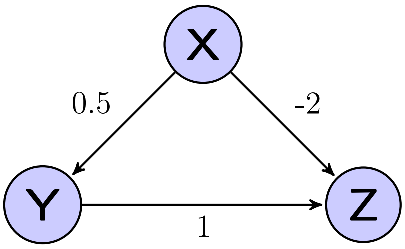


Figure 2.13: Diagram representing the three-variable linear scenario where the mediated causal influence of X on Z via Y exactly cancels out the direct effect, misleadingly indicating independence. The values on the edges indicate the strength of the causal influence.

Definition 2.11.9 (Stability)**.** *A causal model $M = \langle D, \Theta_D\rangle$ is considered to yield a stable distribution when the distribution $P(\langle D, \Theta_D\rangle)$ does not contain any superfluous independencies. Formally, this is the case if and only if the set of conditional independencies in $P(\langle D, \Theta_D\rangle)$, denoted as $I(P(\langle D, \Theta_D\rangle))$, is a subset of $I(P(\langle D, \Theta'_D\rangle))$ for any alternate parameter set Θ'_D.*

For a more intuitive understanding of this definition, refer to Jonas Peters [23].

2.11.3 Structure Learning Algorithms

There are multiple forms of structure learning algorithms [15, 23, 28]. We have now developed sufficient theory to understand the basic ideas behind several classical causal discovery algorithms. In general, there are three broad classes of such algorithms. The first of these are the most natural and make use of several properties we have previously discussed. These are the *constraint-based* methods, some of which we go into detail below. *Score-based* methods are distinct in that they make use of some scoring function/heuristic to evaluate the quality of a given causal structure during the learning phase. Finally, there is a distinct class of *functional causal models* (FCM) approaches. These are methods that use a functional form to evaluate the quality of a given causal structure. This last class of method is less in line with the manner in which we have developed our understanding of causal models. However, it is worth mentioning as it is a popular class of methods in the causal discovery literature.

A special note to the reader: causal discovery algorithms are attempting to do a very difficult and fundamental thing, which is to uncover some underlying hidden nature of the world. This is effectively what we are doing in science in general, with the purpose of the scientific method being to uncover the elusive truth of such models. As such, it is important to understand that every method described here is at the mercy of the assumptions taken to develop the method. For example, one constraint-based method might be very accurate in learning the best Markov-equivalence class given some observational data, but another method might be more useful in practice. That said, for the sake of purity, this thesis has some bias toward methods that are faithful to the underlying data, even if this doesn't always result in the best efficiency of learning or best accuracy in some practical environment.

Constraint-Based Methods

Given no hidden variables, the stability assumption ensures that every distribution has a unique minimal causal structure. This follows from the earlier discussion on the equivalence of causal models given equivalent dependency structures. This is the key that opens up the possibility of algorithmically learning causal structures from observational data. The idea is that since the (unique) equivalence class of minimal causal structures boils down to dependencies, the structures should be learnable via independence tests between variables in the system. What can be feasibly learnt is a graphical structure which has some directed and some undirected edges, representing what is known about the constraints on the possible structure. This semi-directed, learnable graph is called a *pattern*. The IC algorithm, which is an algorithm to identify such patterns, was introduced by Verma and Pearl [29]. We discuss the details of IC, as well as the PC algorithm, in Appendix A.4.

Score-based methods

Transitioning from the intuitive constraint-based algorithms, we now consider *score-based* methods. At their core, these algorithms harness the power of quantitative metrics, or scores, to determine the degree of congruence between the proposed causal model and the observed data. In a sense, every potential causal structure gets assigned a score based on its reflection of the empirical data distribution. The overarching objective is to pinpoint the structure that garners the highest score, suggesting it to be the most probable representation of the causal dynamics.

What distinguishes score-based approaches is their reliance on scoring functions or heuristics to ascertain the quality of a given causal structure during the learning process. While the Bayesian Information Criterion (BIC) is a frequently employed score in this context, it is the principle of decomposability that renders these methods particularly effective. Decomposability ensures that the cumulative score can be broken down into independent scores from individual variables or subsets thereof, making the method feasible and efficient in higher-dimensional contexts.

In our exploration of score-based methods, we will particularly highlight the Greedy-Equivalence Search (GES) algorithm [30] as this is a simple example of how scores can systematically guide the construction and refinement of causal models. We also make use of GES in later experiments and investigations undertaken during this project.

Greedy-Equivalence Search (GES) GES, in contrast to PC, starts with an empty graph. GES works by iteratively adding edges whenever some score (e.g. BIC) is improved by adding that edge. Once all edges have been added, the algorithm works in reverse to remove edges one by one, where the same score is improved by edge removal. This should yield some minimal model class that describes the data. In the large sample limit, GES and PC both converge on the same Markov Equivalence Class under similar assumptions.

- In the forward phase, edges are greedily added to the graph. For each pair of non-adjacent nodes, the algorithm evaluates the impact of introducing a new edge on a given score, such as the Bayesian Information Criterion (BIC). If the score suggests that the model fits the data better with the edge added, the edge is incorporated into the graph. This phase continues until no more edges can be added that improve the score.
- Subsequently, in the backward phase, the algorithm evaluates the impact of removing edges. As in the forward phase, if the score suggests the model is a better fit without a particular edge, that edge is deleted. This process continues until removing further edges no longer improves the score.

The primary allure of GES is its ability to find a minimal causal model that describes the data well, based on the given score. It's worth noting that, under the same assumptions, both GES and PC, in the limit of large samples, are consistent in identifying the same Markov Equivalence Class. This demonstrates the robustness and reliability of these algorithms in the context of causal discovery.

Functional Causal Models (FCM)

Functional Causal Models (FCM) approach causal discovery by representing each variable as a function of its direct causes. This perspective moves away from merely associating variables (as in correlation-based approaches) or assessing conditional independencies (as in constraint-based methods). Instead, it emphasises modelling the

Algorithm 1: Greedy-Equivalence Search (GES) algorithm. Adapted from [30].

```
Input: Dataset $D$ with a set of variables $\mathbf{V}$, scoring function $score(\cdot)$
Output: The undirected graph $G$ with a set of edges $\mathbf{E}$
Initialise an empty graph $G$ with nodes $\mathbf{V}$ and no edges
Forward Phase:
repeat
    for each pair of non-adjacent vertices $X$ and $Y$ in $G$ do
        Compute $score_{add} = score(G \cup \{X, Y\}) - score(G)$
        if $score_{add} > 0$ then
            Add edge $(X, Y)$ to $G$
            Update $\mathbf{E}$ to include $(X, Y)$
until no further increase in score;
Backward Phase:
repeat
    for each edge $(X, Y)$ in $\mathbf{E}$ do
        Compute $score_{remove} = score(G \backslash \{X, Y\}) - score(G)$
        if $score_{remove} > 0$ then
            Remove edge $(X, Y)$ from $G$
            Update $\mathbf{E}$ to exclude $(X, Y)$
until no further increase in score;
```

explicit functional relationships that generate the observed data. FCMs are particularly effective in contexts where the relationships among variables are deterministic, albeit possibly obscured by external noise. FCMs place emphasis on modelling the specific functional mechanisms that underlie the observed interactions, presenting a clear departure from other algorithmic strategies.

Unlike score-based methods, which prioritise models based on their goodness-of-fit to the data using specific scoring criteria, FCMs seek to unearth the actual functional form of the causal relationships. Similarly, while constraint-based methods use statistical tests to identify conditional independencies among variables and subsequently construct a causal graph, FCMs begin with the assumption that a clear deterministic functional relationship exists and strive to decipher it.

The Linear non-Gaussian Acyclic Model (LiNGAM) is an example of an FCM. It assumes that each observed variable is a linear function of its direct causes and some non-Gaussian noise, allowing it to uniquely identify causal structures under certain assumptions. We provide the details of this approach in Appendix A.4.

Generative Approaches

Generative methods in causal discovery emphasise modelling the process by which data are generated. This is in stark contrast to classical causal discovery, which typically focuses on identifying patterns or structures in already observed data. Generative approaches operate on the principle that if we can closely mimic the process by which data are produced, we can then make meaningful interventions in the model and observe the outcomes, thus inferring causality.

The recent upsurge in the application of deep learning techniques has made it increasingly feasible to model high-dimensional data distributions. Generative models, such as Generative Adversarial Networks (GAN) [31] and Variational Autoencoders (VAE) [32], have been at the forefront of recent advances in generative AI applications. In the context of causal discovery, these models offer the potential to approximate the data-generating process with high fidelity, even when that process is intricate or non-linear [33].

Deep learning's ability to model non-linear and intricate relationships is particularly beneficial for causal discovery. When combined with generative modelling, we have tools that can simulate potential interventions, even if such interventions have never been observed in the real world. This novel approach stands in contrast to classical methods, which rely heavily on observed data and statistical tests. Some deep causal models leverage these generative capabilities. We provide an example of how one might employ a GAN for causal discovery in Appendix A.4.

Regarding scoring and evaluation of learnt graphs, various graph distance evaluation metrics are used. We refer the reader to Appendix A.5.1 for a brief discussion.

We further refer the reader to Appendix A.6 for a discussion about others methods that hint at causal dynamics and are common in the field of statistical learning.

2.12 Conclusion

This chapter provided an overview of the core concepts and methodologies required to address the causal aspects of the research questions outlined in this thesis. A key distinction was made between causal *inference* and causal *learning*, emphasising their respective roles in understanding causal relationships.

Graphical methods in causality were then introduced and discussed. These methods have been highlighted due to their utility in visualising and modelling causal relationships between variables. Their use aids in both theoretical and empirical analyses.

The foundational knowledge from this chapter will be used to guide and inform the methodologies and analyses in subsequent chapters, ensuring a structured and systematic approach to answering the research questions.

Chapter 3

Intelligence and Learning Agents

This chapter moves away from foundational causality theory to the study of intelligent agents, tackling the challenge of defining and quantifying intelligence. We offer a brief perspective and suggest key resources for further reading. The chapter then covers crucial aspects of deep learning and reinforcement learning (RL), focusing on the core models that form the basis of RL and related fields. It connects these models to causality, leading into the formal concepts of multi-agent systems. By the end of the chapter, readers will have a basic understanding of the relationship between causality and RL, a theme explored in later chapters. Throughout this chapter we introduce content that is particularly relevant for addressing the research questions, with the theme of *action as intervention*.

3.1 Introduction

The idea of creating machines capable of learning from experience traces back to the early days of computer science. In 1947, Alan Turing delivered a lecture to the London Mathematical Society [34], envisioning a machine with the ability to modify its initial set of instructions. Drawing parallels with a student acquiring knowledge, Turing stated,

"*It would be like a pupil who had learnt much from his master, but had added much more by his own work. When this happens I feel that one is obliged to regard the machine as showing intelligence.*" - Alan Turing, 20 Feb 1947.

The pursuit of machine intelligence raises the challenge of defining what 'intelligence' entails. This act of defining often alters the very attribute we attempt to measure, reminiscent of Goodhart's Law [35]. As we progress in AI research, the yardstick for machine intelligence continually evolves, evident from Turing's *Imitation Game* [36], to the triumphs of Deep Blue in chess [37], and DeepMind's achievements in Go [38]. Despite these advancements, determining whether such systems exemplify *intelligence* remains an enigma.

Transitioning from this broad view, let's consider RL, a sub-field that sits at the intersection of optimal control, animal psychology, and artificial intelligence, among other areas [39]. The foundational principles of RL draw inspiration from the study of animal behaviour, encapsulated in Thorndike's 1898 *Law of Effect* [40]. Modern RL has made significant strides, with various RL algorithms showcasing unparalleled performances across benchmarks [38, 41–45]. In particular, the application of RL to fine-tune large language models (LLM) towards human preferences has garnered immense interest [46, 47].

In this context, we aim to define intelligence, a crucial step in identifying gaps in current methods and encouraging new interdisciplinary approaches, including insights from causality. The chapter discusses intelligence, leading into key topics of RL, multi-agent dynamics, and the vital role of causality. While offering a broad overview, the chapter refers readers to additional resources for details not central to the thesis's main argument.

3.2 Measures of Intelligence

"Intelligence is the efficiency in which you turn experience into generalisable programs." - Francois Chollot, 2020 [48].

Following this definition, a long-standing criticism of AI research is that it has produced predominantly narrow and task-specific systems. Chollet [48] suggests that this is largely due to the fact that, as researchers, we've historically excelled at defining objectives in a narrow, measurable, and actionable way. The intuitive, common-sense understanding of intelligence often falls short when trying to establish a concrete objective for developing intelligent systems. To emphasise the challenge, Chollet mentions that in 2007 there were a "minimum of 70 definitions of intelligence in related literature". A recurring theme among these definitions is the idea that "intelligence measures an agent's ability to achieve goals in a wide range of environments."

Before examining the latest models and formulations of learning systems, it's crucial to deeply understand these definitions, particularly since many modern approaches implicitly aim for emulating intelligent (human-level) behaviour. This ambition can be seen as early as Marvin Minsky's definition of AI [49], who is quoted as saying "AI is the science of making machines capable of performing tasks that would require intelligence if done by men."

Central to these discussions is the concept of intelligence as *general learning ability*, or *generalisability*. Two main challenges arise from this perspective: (1) how to design tasks that effectively measure skill, and, in turn, (2) how to use skill proficiency as a proxy for gauging intelligence. It's essential to recognise the implicit assumption that intelligence should entail a degree of memory efficiency. Relying solely on memorisation wouldn't truly capture the essence of intelligence but would simply demonstrate recall. This often becomes the crux of debates centered around the claim that "neural networks are all you need." Both lookup tables and neural networks (NN) serve as methods to represent functions. Their key difference lies in how efficiently they capture and recreate the underlying data. To put it plainly, NNs are better equipped to handle scalability, especially with larger datasets.

In the quest for general intelligence, benchmarking against human intelligence seems a logical step. Although there are various ways to define and measure intelligence, Chollet introduces the *G factor* as one human-centric metric. This statement suggests that the highest form of intellectual reasoning is achieved through extreme generalisation. Figure 3.1, extracted from the original paper, delineates the relationship between general intelligence and both broad and local generalisation, situating them within a hierarchy of cognitive abilities.

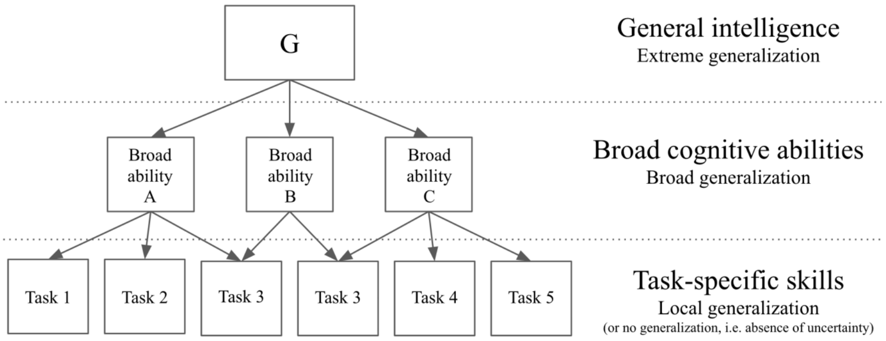


Figure 3.1: Hierarchical model extracted from [48] depicting cognitive abilities in the context of generalisation.

Venturing a bit off the main path, it's intriguing how Chollet's ideas resonate with Pearl's perspectives on causality [15]. It is important to note, however, that this concept of causality being crucial to human reasoning is not a novel idea, with its roots tracing back to early philosophical discourses [50]. While Pearl doesn't refute that certain micro-phenomena in physics challenge traditional notions of deterministic cause-and-effect, he believes that causal modelling remains an integral framework for human reasoning. Chollet argues that for a machine to be considered genuinely intelligent, it should be endowed with a set of core knowledge priors. There's a vibrant discussion on how these priors correlate with the explicit assumptions in Pearl's graphical models. These insights set the stage for our subsequent deep dive into RL theory, culminating in a thorough exploration of its interplay with causal frameworks.

3.3 Exploring Deep Learning in Machine Learning

In the field of machine learning (ML), the integration of deep learning methods represents a notable advancement, particularly in their ability to handle complex functions and discern patterns in large datasets [51]. This section considers how deep learning uses artificial NNs to create advanced systems that exhibit intelligent behaviour.

Deep learning, a branch of ML, is increasingly used for function approximation, thereby improving performance metrics. This trend is significant for two key reasons: it is vital to understand why deep learning techniques work effectively, and it is important to recognise the situations where they are most advantageous. Consequently, this section focuses on elucidating the basic concepts of deep learning, starting with an introduction to NNs, followed by placing these networks in the context of graphical methods and other function approximation techniques. It is worth noting that this section will not discuss the mathematical details of deep learning but will highlight useful resources in this area.

An artificial NN (ANN, or NN) is a computational system that mimics biological NNs, consisting of nodes or *neurons* linked by *synapses* with variable weights, forming a weighted DAG [52–54]. This structure enables the representation and learning of complex functions, supported by modern computational frameworks. NNs' ability to represent various functions is supported by universal approximation theorems, which apply to networks with different depths and widths [55, 56]. Additionally, ongoing research is exploring questions about the capabilities and universality of graph NNs [57, 58].

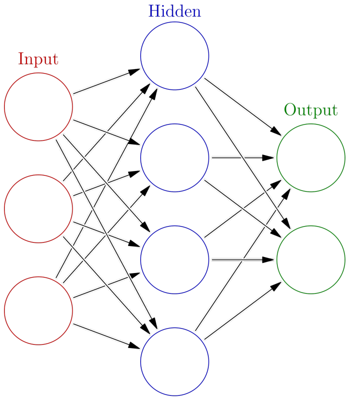


Figure 3.2: A basic NN with 3 input neurons (red), 1 hidden layer of 4 neurons (blue), and 2 output neurons (green).

Theorem 3.3.1 (Deep Neural Networks and Universal Approximation)**.** *A single-layer feed-forward network, with a finite number of neurons in its hidden layer, can approximate continuous real-valued functions on compact domains in* $\mathbb{R}^n$ *to any desired accuracy.*

This theorem highlights the capacity of NNs to represent a wide range of functions by adjusting parameters and weights. It suggests that under certain conditions, NNs can approximate any continuous function, which is both theoretically interesting and practically relevant in fields like RL and various ML applications.

Deep neural networks (DNN) not only represent a variety of functions but also enable the learning of approximations that accurately map inputs to outputs. A key component of this learning process is the *backpropagation* algorithm, which is described in Algorithm 2 [59].

Backpropagation is an optimisation technique that adjusts a network's weights and biases to minimise prediction errors. This optimisation involves calculating the gradient of the cost function with respect to each parameter, using matrix multiplication for efficiency. Understanding the link between this method and the universal approximation theorem is important for grasping how changes in functions are managed. The backpropagation process and its associated notation are detailed below:

- **Training Data**: Defined as $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, where each element comprises an input x and its corresponding target output y.
- **Network Parameters**: Represented by W for weights and b for biases, these parameters are adjustable within the NN. Starting with random values helps prevent symmetric convergence issues.
- **Training Iterations**: The training process involves multiple epochs, each refining the model until certain criteria, like convergence, are met.

- **Forward Propagation**: Beginning with $a^{(1)} = x$, each layer's activations are calculated using $a^{(l)} = \sigma(W^{(l-1)}a^{(l-1)} + b^{(l)})$, where σ represents the activation function.
- **Error Measurement**: The loss is calculated as $J = \text{loss}(a^{(L)}, y)$, indicating the difference between the network's output, $a^{(L)}$, and the actual target, y.
- **Backward Propagation**: Starting from the error in the output layer $\delta^{(L)}$, the error terms for preceding layers are computed and used for updating the weights and biases, influenced by a learning rate α.
- **Element-wise Multiplication**: Denoted by $\odot$, this represents the Hadamard product, used for element-wise operations.

Algorithm 2: Algorithm for Backpropagation in Neural Network Training

```
Data: Training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, learning rate $\alpha$
Result: Refined weights and biases in the neural network
Initialise weights $W$ and biases $b$ randomly;
while not converged (e.g., based on loss threshold or max iterations) do
    for each $(x, y)$ in training set do
        // Forward propagation
        $a^{(1)} = x$;
        for $l = 2$ to $L$ do
            $z^{(l)} = W^{(l-1)}a^{(l-1)} + b^{(l)}$;
            $a^{(l)} = \sigma(z^{(l)})$;
        // Loss calculation for the chosen loss function
        $J = \text{loss}(a^{(L)}, y)$;
        // Backward propagation
        Initialise gradients of weights and biases to zero;
        $\delta^{(L)} = \nabla_a J \odot \sigma'(z^{(L)})$;
        for $l = L - 1$ down to $2$ do
            $\delta^{(l)} = (W^{(l)^T}\delta^{(l+1)}) \odot \sigma'(z^{(l)})$;
            $W^{(l)} \leftarrow W^{(l)} - \alpha\delta^{(l)}(a^{(l)})^T$;
            $b^{(l)} \leftarrow b^{(l)} - \alpha\delta^{(l)}$;
```

Through iterative refinements directed by backpropagation, NNs progressively learn optimal mappings, making them versatile in tasks from regression to classification.

While NNs were initially inspired by the biological brain, with a goal to mimic its functions [60], contemporary research has shifted towards tailoring network designs for specific applications. For instance, convolutional neural networks (CNN), which excel in object detection and image classification, employ a series of *convolutions* to extract and emphasise image features. CNNs have become indispensable in visual RL tasks. Moreover, advances in recurrent neural networks (RNN) [61], capable of processing sequential data like speech or text, represent another significant stride. RNNs, and specifically long short-term memory (LSTM) models [62], address challenges such as the vanishing gradient problem, enabling the processing of long-term dependencies in data.

3.4 Causality and Neural Networks

NNs, as supported by the universal approximation theorems, have the innate ability to approximate almost any function, given an appropriate architecture and ample data. They adjust their weights to closely emulate a target function, driven by error minimisation. This adaptability lends them utility in causal inference, especially when modelling relationships in causal DAGs.

While these networks can capture complicated relationships, it's important to understand their inherent limitations concerning causality. The training of ML methods that use NNs for function approximation are particularly good at identifying correlations within static data (i.e. statistical learning) but often struggle with distinguishing genuine causal relationships from spurious correlations (causal learning). The fundamental challenge lies in their observational nature. A NN trained on static data captures prevalent patterns but may not be equipped to predict outcomes under novel interventions, primarily because it lacks insights from data that embodies these interventions.

To illustrate, a NN might grasp the association "if A, then B" from observational data. However, when confronted with the task of predicting the outcomes after actively intervening on A, it might not perform adequately if such interventions weren't part of its training data. This brings us to the domain of RL - interventions. Unlike traditional supervised and unsupervised methods, RL is framed as an agent taking actions, and can thus engage with the repercussions of its own actions, making it a promising avenue for understanding and predicting the outcomes of interventions. In the next chapter we will investigate the shared domain of causality and RL, highlighting the potential of causal RL to bridge the gaps we've identified. This will allow us to begin tackling the research questions. That said, we first need to develop RL theory.

3.5 Reinforcement Learning

RL is a distinct branch of ML that differs from both supervised and unsupervised learning paradigms. Supervised learning relies on labeled data to make predictions, while unsupervised learning seeks patterns in unlabeled data. In contrast, RL involves agents that interact with environments, aiming to determine optimal actions based on received rewards and consequences of their actions.

Traditionally, RL operates on the principle of trial-and-error learning. An agent, situated in a state within an environment, makes decisions or takes actions, eliciting responses from the environment, typically in the form of rewards. The agent's ultimate quest is to maximise the cumulative reward, termed the *return*, G_t. However, the agent often grapples with limited visibility of the environment's inner workings, reminiscent of the fog of war in strategic games such as Montezuma's Revenge[1]. Here, the agent must deftly navigate and decide based on imperfect information.

As touched on earlier, intuitively understanding the nature of RL can be enriched through the lens of causality. Schölkopf [63] considers this, arguing that an agent's actions can be seen as interventions in the environment, altering its trajectory. The reactions to these interventions, coupled with rewards, illuminate the cause-and-effect

[1] Montezuma's Revenge serves as a benchmark for RL agents, demanding both strategic exploration and skill mastery for successful gameplay.

relationships within the environment. The agent's task is akin to a scientist determining causal relations, albeit in a digital landscape where the consequences of actions might not always be immediately evident.

Recent developments in RL have further enriched its taxonomy, introducing divisions like supervised RL and unsupervised RL. In supervised RL, the rewards, which serve as learning signals for the agent, are explicitly defined. Meanwhile, unsupervised RL sees agents propelled by intrinsic motivations, like curiosity, carving their own learning path.

Delving deeper, RL algorithms can be broadly categorised into two paradigms: model-free RL (MFRL) and model-based RL (MBRL). While MFRL agents learn directly from their experiences, honing policies and value functions, MBRL agents take a more analytical approach. They strive to 'understand' and model the environment's dynamics, thereby indirectly deducing optimal actions. Though MFRL has made waves in domains like robotics and video games, it often falters in terms of sample efficiency, especially in environments mirroring the multifaceted dynamics of the real world. In contrast, MBRL, with its capability to distil the environment's complexity, can significantly enhance both efficiency and performance. This, of course, related directly to **RQ2**: *Does a causal model improve the sample efficiency and/or coordination of learning agents in a decentralised learning task?* However, MBRL is not without its pitfalls, especially when modelling inaccuracies creep in, skewing the agent's understanding and leading to erroneous decisions. This intriguing interplay between the two paradigms, especially the emphasis on modelling in MBRL, draws parallels with the model-centric approach in causal inference, setting the stage for deeper explorations in subsequent sections. We touch on this in more depth in Section 3.9.

Table 3.1: Different Divisions and Paradigms in Reinforcement Learning

RL Classification	**Description**
Supervised RL	Rewards are explicitly defined, serving as learning signals for the agent.
Unsupervised RL	Agents are driven by intrinsic motivations such as curiosity, defining their own learning path.
Model-Free RL	Agents learn policies and value functions directly from their interactions with the environment. Commonly used in robotics and video games, but may struggle with sample efficiency in complex environments.
Model-Based RL	Agents attempt to understand and model the environment's dynamics, deriving optimal actions indirectly. Though potentially more efficient and high-performing, inaccuracies in the model can lead to poor decisions.

3.5.1 The Reinforcement Learning Problem

Simply put, RL is about finding the best way to choose actions based on certain situations to get the most reward over time [64]. Imagine an RL 'agent' that moves around in a changing environment. The way this environment changes depends on the agent's actions. The agent tries different actions, learns from its experiences, and aims to figure out the best set of actions to take.

To systematically tackle this problem, we use the Markov Decision Process (MDP). Think of the MDP (Definition 3.5.3) as a set of rules or a framework that's well-suited for basic RL problems. Many traditional RL methods use a tool called dynamic programming. This tool works best when we have a complete model — a model that can perfectly predict what will happen for any action in any situation. It's like how a video game designer, using a game physics engine, can predict what will happen in the game based on certain actions.

We can think of RL as a way of guessing dynamic programming. This guessing or "approximation" is what makes some areas of RL unique. In this section, we'll dive deeper into these ideas, preparing the ground for a detailed discussion on key RL methods. We'll start by explaining a typical MDP problem.

Example 5 (Taxi Environment)**.** *The Taxi environment, as described by [65], provides an illustrative introduction to RL scenarios. Envision a discrete 5x5 grid world, representing a simplified cityscape. Within this environment, a taxi driver navigates to pick up a passenger and subsequently drops them off at specified locations. While, in a more complex representation, various factors such as fuel consumption, toll roads, or passenger comfort might influence the driver's decisions, we'll streamline the scenario. The primary goal is to minimise the total discrete steps taken to collect and deliver the passenger. Each step incurs a cost of 1, with a positive reward presented upon successful drop-off.*

This environment features four distinct pick-up and drop-off locations, colour-coded as Red, Green, Yellow, and Blue. At the commencement of an episode, both the taxi's and passenger's starting positions are randomised. The taxi's objective then is twofold: to locate and pick up the passenger and to deliver them to the predetermined colour-coded destination. Each episode concludes upon successful delivery. It's worth noting the penalty system embedded within the game mechanics. Negative rewards are imposed for any erroneous pick-up or drop-off attempts, alongside the cost per step taken.

The ensuing discussions will harness this Taxi environment as a conceptual aid, extrapolating various RL methodologies and their potential ramifications.

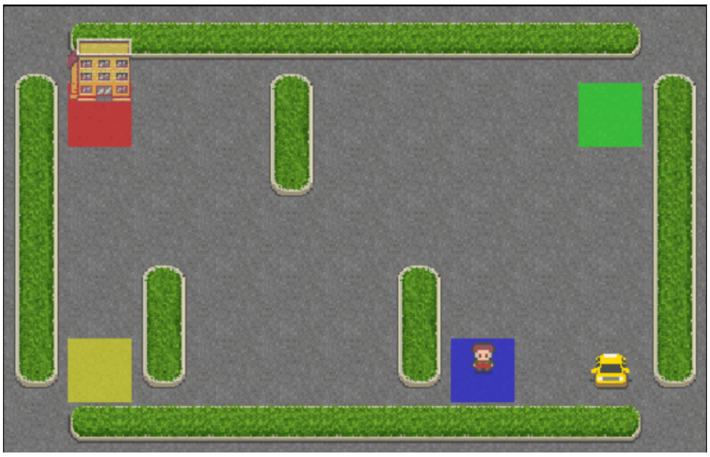


Figure 3.3: A visual depiction of the Taxi environment [65].

The Agent-Environment Interface characterises the iterative interaction between an agent and its environment, which in turn provides essential data that facilitates agent learning based on environmental responses. This interaction is schematically presented in Figure 3.4. At any given time-step $T = t$, the environment is characterised by a specific state s_t. With this state as context, the agent determines an action a_t. The execution of this action leads the environment to transition into a new state s_{t+1}, which subsequently yields a corresponding reward r_t. Conceptually, each interaction cycle produces a data tuple, encapsulating state, action, reward, and the subsequent state.

In the Taxi scenario, a practical example can help illustrate these abstract concepts. Here, s_t represents the current positions of both the taxi and the passenger. The action a_t corresponds to the decision made by the taxi driver, which can include several possibilities: moving south (0), north (1), east (2), or west (3); picking up a passenger (4); or dropping off a passenger (5). The reward r_t reflects the rewards or penalties obtained by the taxi driver at each time-step t, which are largely dependent on the action executed. Finally, s_{t+1} denotes the subsequent positions of the taxi and passenger following the action.

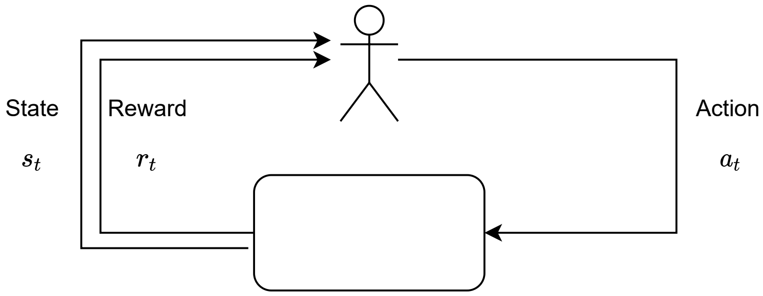


Figure 3.4: Schematic representation of data flow within a MDP. An agent, based on a given state, chooses action A_t. Subsequent to this choice, the environment reveals the following state alongside the associated reward. This dynamic continually iterates, shaping the agent's learning trajectory [64].

3.5.2 Formalising the RL Problem

The RL problem can be conceptualised by recognising that the rewards an agent receives are influenced by its current state and the actions taken. In such scenarios, the historical sequence leading to the present state is typically irrelevant for the immediate future, characterising what is known as the Markov property. Though there exist RL environments that remember past actions of the agent [66], these are exceptions to the standard RL framework and are beyond the scope of this discussion. Nonetheless, non-Markovian problems in RL can be tackled using memory-based structures or approximations with an MDP [67, 68].

Definition 3.5.1 (Markov Property)**.** *A state is said to have the Markov property if, for every time step* $t = 1, 2, \ldots,$ *the conditional probability*

$$P(X_{t+1} = i_{t+1} | X_t = i_t, \ldots, X_1 = i_1, X_0 = i_0) = P(X_{t+1} = i_{t+1} | X_t = i_t).$$

In this framework, agents operate within environments where each state adheres to the Markov property. A *Markov chain* is a discrete-time stochastic process that arises when this property is universally applicable across states, indicating a sequential dependency among them. This concept is encapsulated in a tuple (S, P), which includes the set of states S and their transition probabilities P.

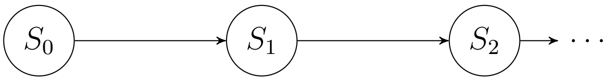


Figure 3.5: A causal DAG illustrating the Markov property of a Markov chain. Here S represents states of the environment in the context of RL.

In an RL setting, agent-environment interactions result in rewards, leading to an augmentation of the Markov chain concept, resulting in a Markov reward process. This extension involves rewards at various time steps and introduces the *discount factor*, γ, allowing the agent to value future rewards appropriately.

Definition 3.5.2 (Markov Reward Process)**.** *A Markov reward process (MRP) adds a reward dimension to the structure of a Markov chain. It is formally defined as a 4-tuple* (S, P, R, γ)*, where* S *denotes states,* P *transition probabilities,* R *rewards, and* γ *the discount factor.*

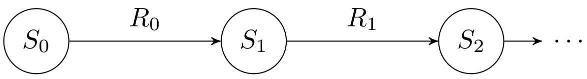


Figure 3.6: Illustration of a Markov reward process (MRP) with states S and rewards R associated with transitions between states.

This structure highlights the dependency on states that is fundamental to RL problems, differentiating them from scenarios like multi-armed bandits, which do not rely on previous events. However, this model still does not encompass the agent's decision-making processes. The incorporation of actions leads to the more comprehensive Markov Decision Process.

Definition 3.5.3 (Markov Decision Process (MDP))**.** *A Markov Decision Process (MDP) is a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. It is defined as a 5-tuple* (S, A, P, R, γ)*, where* S *represents a finite set of states in the environment,* A *denotes a finite set of actions available to the agent,* $P : S \times A \times S \to [0, 1]$ *is the transition probability function that specifies the likelihood of transitioning from one state to another given an action,* $R : S \times A \times S \to \mathbb{R}$ *is the reward function that assigns a numerical reward to each transition between states due to an action, and* γ *is a discount factor in the range* $[0, 1]$ *that weighs the importance of future rewards versus immediate rewards.*

Adding decision-making capabilities necessitates an understanding of how actions are selected, which forms the crux of the RL problem. In RL, an agent's actions are guided by a *policy*. A deterministic policy can be defined as a function $\pi : S \to A$, assigning a specific action a to each state s. In contrast, a stochastic policy provides a probability distribution over actions for each state, represented as:

$$\pi(a|s) : S \times A \to [0, 1]. \tag{3.1}$$

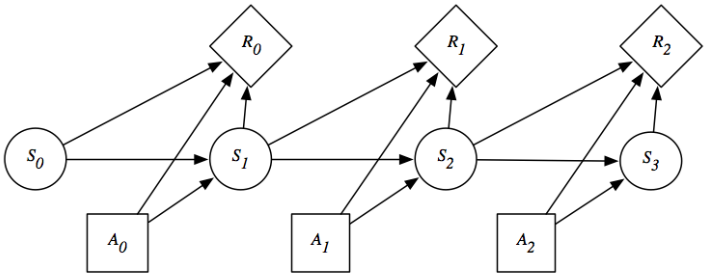


Figure 3.7: Figure showing a finite part of a Markov decision process. Here states, actions, and rewards are labelled. Transitions are represented by edges. Figure taken from [69].

In this representation, $\pi(a|s)$ indicates the probability of taking action a in state s. Often, policies are parameterised by a set θ, denoted by π_θ, which allows for policy optimisation, a topic further explored in section 3.8.

With this added complexity, new challenges emerge, such as those in certain computer games where the agent has only partial or altered access to the state space. This leads to the concept of a partially observed Markov decision process (POMDP) [70].

Definition 3.5.4 (Partially Observed Markov Decision Process (POMDP))**.** *A Partially Observed Markov Decision Process (POMDP) extends an MDP to account for situations where the agent does not have complete information about the current state. Instead, the agent receives observations that provide partial information about the state. Formally, a POMDP is expressed as a 7-tuple* $(S, A, P, R, \Omega, O, \gamma)$*, where:*

- *S, A, P, R, and γ are as defined in the MDP.*
- *Ω is a set of observations, representing all the possible observations an agent can receive about the current state.*
- *O is an observation probability function, $O(o|s', a)$, that gives the probability of receiving observation o after taking action a and transitioning to state s'.*

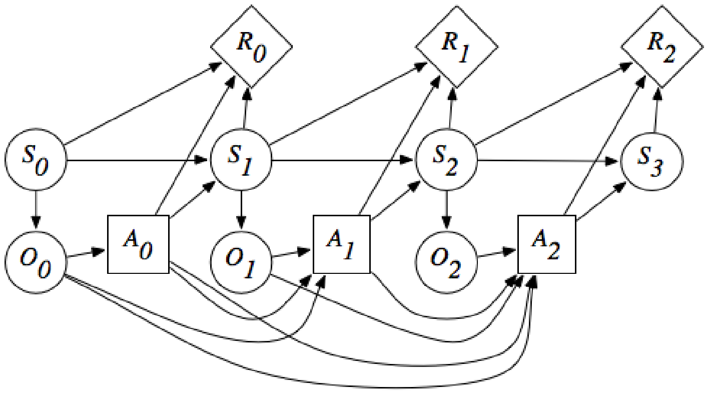


Figure 3.8: Figure showing a finite part of a partially observable Markov decision process. Here states, observations, actions, and rewards are labelled. Transitions are represented by edges. Figure taken from [69].

This modification reflects the reality that partial observability disrupts the clear link between an observation and the environment's true state. Agents cannot directly correlate observations with the ideal Markov states, which ideally provide comprehensive information about the current state of the environment. The subsequent sections will focus primarily on fully observable MDPs, acknowledging that many real-world situations often involve partial observability. The assumption is that, in many cases, the MDP framework and the algorithms developed for it are sufficiently robust.

3.6 Optimality in Reinforcement Learning

Given the context of RL, an important question arises: given the structure of agent-environment interactions, how can agents consistently make decisions that maximise their cumulative rewards? This pursuit towards the best course of action forms the essence of the concept of optimality, which, in terms of the RL formulation, is core to emergent 'intelligent behaviour' – at least philosophically [71]. Fundamentally, RL aims to maximise cumulative rewards through the appropriate selection of actions. This objective can be represented mathematically via a value function, which quantifies the anticipated cumulative reward an agent can accrue from a given state or under a specific policy. Deriving the most advantageous policy, one that optimises this value function, epitomises an optimisation challenge.

Optimisation is an important research area of mathematics and machine learning, especially within the context of deep learning. The process of "training" or "learning" frequently involves the iterative refinement of model parameters to enhance predictive accuracy. Gradient descent (or similar methods) is foundational to highly performant training of dense NNs. By utilising the gradient of the loss function, model parameters can be systematically adjusted to minimise prediction errors. Various optimisation schemes have become popular in machine learning research and applications. The specifics of this are outside the scope of the core argument of this thesis, but we provide a brief discussion in Appendix B.2.

3.7 Value-based Methods

In RL, agents are tasked with maximising their performance within defined environments. When framing these RL challenges as MDPs, a promising approach is to assign a 'value' to each state within the environment. A clear analogy for this concept can be drawn from the world of chess. In chess, players evaluate board positions relative to their adversaries, much like how RL agents assess the expected cumulative rewards from specific states [37, 72, 73]. By continually seeking states of higher value, agents can be steered towards more favourable outcomes, and, ultimately, optimal behaviours. This idea forms the basis for *value-based methods* — a category of RL algorithms that emphasise optimising state and state-action values.

3.7.1 Value and Value Functions

In RL, the predicted 'value' of a state provides a glimpse into its potential worth in terms of forecasted future rewards.

Example 6 (Value: A Chess Analogy)**.** *To illustrate, consider the game of chess. Different players, based on their expertise, may perceive the value of a board configuration distinctively. While a novice might focus on immediate threats, a Grand Master looks deeper. They can discern hidden strategic avenues in board states that may seem un-*

inviting to lesser-skilled players. However, their skill isn't simply about evaluating more lines or delving deeper into possible moves. The vast quantity of possible chess continuations ensures that even Grand Masters touch upon only a fragment of all sequences. Their true strength lies in recognising patterns and relying on a mental repertoire of positions and sequences acquired over years. This pattern recognition enables them to swiftly identify promising strategies, thereby effectively navigating through an array of opponent tactics.

The concept of a state's value in RL is intrinsically linked to the policy employed by the agent. Denoted by $v_\pi(s)$, it reflects the expected cumulative rewards starting from state s under policy π.

The *state-value function* for a policy π, $v_\pi(s)$, represents the expected total rewards initiated from state s and following policy π. Mathematically, it is expressed as:

$$v_\pi(s) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s\right] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right], \text{ for all } s \in \mathcal{S}.$$

Similarly, the *state-action value function*, $q_\pi(s,a)$, measures expected rewards when an agent, starting from state s and taking action a, continues to follow policy π. This is defined as:

$$\begin{aligned} q_\pi(s,a) &\doteq \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a\right] \\ &= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right], \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}. \end{aligned}$$

While $v_\pi(s)$ assesses the expected return from state s under policy π without specifying an action, $q_\pi(s,a)$ evaluates the impact of a particular action a taken from state s. When the best possible action at state s is chosen according to policy π, these two values converge, leading to $v_\pi(s) = \max_a q_\pi(s,a)$, which is characteristic of an optimal value function.

Furthermore, the Advantage function, denoted by $A_\pi(s,a)$, evaluates the relative merit of an action a compared to the average action at state s under policy π. It is defined as the difference between the state-action value function and the state-value function:

$$A_\pi(s,a) = q_\pi(s,a) - v_\pi(s)$$

This metric is crucial for identifying actions that are particularly effective or ineffective in a given state, thereby guiding the agent to refine its policy for improved decision-making over time.

3.7.2 Bellman Equations

The Bellman equations [74], derived from the recursive definitions of value functions, form the backbone of RL. They express the recursive relationship between a state's value and the expected values of its successor states when following a policy π.

Given the value function expansion:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v_\pi(s')\right],$$

the state-action value function, $q_\pi(s,a)$, can be expressed as:

$$q_\pi(s,a) = \sum_{s',r} p(s',r \mid s,a) \left[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s',a') \right].$$

The Bellman equations not only facilitate explicit look-ahead to compute future value but also allow backward propagation of value from future states to the present. This retroactive value propagation forms the crux of the *Bellman optimality equations*, which aim to identify the optimal value functions, $v_\star(s)$ and $q_\star(s,a)$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The relationships between optimal value functions are given by:

$$v_\star(s) = \max_{a\in\mathcal{A}} \sum_{s',r} p(s',r \mid s,a) \left[r + \gamma v_\star(s') \right], \tag{3.2}$$

$$q_\star(s,a) = \sum_{s',r} p(s',r \mid s,a) \left[r + \gamma \max_{a'} q_\star(s',a') \right]. \tag{3.3}$$

For finite MDPs, these Bellman equations can be explicitly solved. For instance, in problems like the Taxi grid-world (Figure 3.3), an exact solution can be obtained with computational complexity $\mathcal{O}(|\mathcal{S}|^3)$. However, as the problem size increases, this becomes computationally challenging. Sutton and Barto [64] point out practical challenges: (1) the need for accurate environment dynamics, (2) computational resource constraints, and (3) the assumption that the Markov property holds.

3.7.3 Value Iteration & Dynamic Programming

Dynamic programming is a method to solve problems by recursively solving and storing solutions of sub-problems. Within the scope of RL and MDPs, dynamic programming techniques are employed to solve the Bellman equations iteratively. The recursive nature of these equations, specifically the Bellman optimality equation, can be converted into an update rule for value functions, facilitating an approach known as *iterative policy evaluation.*

The update rule at time-step k for the state-value function is:

$$v_{k+1}(s) = \sum_{a} \pi(a \mid s) \sum_{s',r} p\left(s',r \mid s,a\right) \left[r + \gamma v_k\left(s'\right) \right].$$

The classical value iteration algorithm, assuming discrete state and action spaces, is as follows:

1. Initialise $v_0(s) = 0$ for all $s \in \mathcal{S}$.
2. For $k = 0, 1, 2, \ldots$:
   (a) Update: $v_{k+1}(s) \leftarrow \max_{a\in\mathcal{A}} \sum_{s',r} p(s',r \mid s,a) \left[r + \gamma v_k(s') \right]$ for all $s \in \mathcal{S}$.
   (b) Terminate if $\max_{s\in\mathcal{S}} |v_{k+1}(s) - v_k(s)| < \theta$.

Given this, the deterministic policy is defined as:

$$\pi'(a_t|s_t) = \begin{cases} 1 & \text{if } a_t = \arg\max_{a_t} A^\pi(s_t,a_t) \\ 0 & \text{otherwise} \end{cases}$$

With bootstrapping, the agent updates its value as:

$$V^{\pi}(s) \leftarrow r(s, \pi(s)) + \gamma\, \mathbb{E}_{s' \sim p(s'|s,\pi(s))}[V^{\pi}(s')].$$

Sutton and Barto [64, pg 78] introduced the *policy improvement theorem*, guaranteeing that an improved policy always yields better or equal value.

Theorem 3.7.1 (Policy Improvement Theorem)**.** *For deterministic policies π and π', if $Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s) \quad \forall s \in S$, then π' is at least as good as π.*

Proof.

$$\begin{aligned} v_{\pi}(s) &\leq q_{\pi}(s, \pi'(s)) \\ &= \mathbb{E}\left[R_{t+1} + \gamma v_{\pi}(S_{t+1}) | S_t = s, A_t = \pi'(s)\right] \\ &= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma v_{\pi}(S_{t+1}) | S_t = s\right] \\ &\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma q_{\pi}(S_{t+1}, \pi'(S_{t+1})) | S_t = s\right] \\ &= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma\, \mathbb{E}_{\pi'}\left[R_{t+2} + \gamma v_{\pi}(S_{t+2})\right] | S_t = s\right] \\ &\vdots \\ &\leq \mathbb{E}_{\pi'}\big[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots | S_t = s\big] \\ &= v_{\pi'}(s) \end{aligned}$$

□

The iterative application of the policy improvement theorem ensures convergence to the optimal policy and value function:

1. Evaluate the value function.
2. Update: $\pi \longleftarrow \pi'$.

Direct value function improvement, without explicit policy evaluation, is the foundational idea behind *value iteration* algorithms. Value Iteration is a dynamic programming method utilised in RL to find the optimal value function for a given MDP. The algorithm iteratively updates the value function until convergence, ultimately allowing the derivation of an optimal policy. The underlying principle is to solve the Bellman optimality equation iteratively for each state.

While the value iteration algorithm iteratively refines the value function until it converges to an optimal value function, another approach is to iteratively improve the policy itself, hence the name, *policy iteration.* Policy iteration consists of two main phases: *policy evaluation*, where the value of a given policy is computed, and *policy improvement*, where the policy is enhanced based on the current value function estimate. This process starts by evaluating a random policy to estimate its corresponding value function. Once the value function for a policy has been determined, it can be used to derive a better policy. These two steps are alternated iteratively until the policy stabilises, indicating it has converged to an optimal policy.

The significant advantage of policy iteration over value iteration is that it often converges to the optimal policy in fewer iterations. However, each iteration can be computationally more intensive due to the need to solve the entire policy evaluation step.

Algorithm 3: Value Iteration algorithm for RL.

```
Input: MDP defined by states $\mathcal{S}$, actions $\mathcal{A}$, transition probabilities
       $p(s', r|s, a)$, rewards $R$, discount factor $\gamma$, and a small threshold $\theta$ for
       convergence.
Output: Optimal value function $v_\star$ and the optimal policy $\pi_\star$
Initialise $v(s) = 0$ for all $s \in \mathcal{S}$
repeat
    $\Delta \leftarrow 0$
    for each state $s \in \mathcal{S}$ do
        $v_{\text{temp}} \leftarrow v(s)$
        $v(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s', r} p(s', r|s, a) \left[r + \gamma v(s')\right]$
        $\Delta \leftarrow \max(\Delta, |v_{\text{temp}} - v(s)|)$
until convergence;
until $\Delta < \theta$
for each state $s \in \mathcal{S}$ do
    $\pi_\star(s) \leftarrow \arg\max_{a \in \mathcal{A}} \sum_{s', r} p(s', r|s, a) \left[r + \gamma v(s')\right]$
```

As we discuss more RL techniques, we encounter methods that blur the lines between policy and value-based approaches. A quintessential example is *Q-learning*, which directly estimates the optimal action-value function without needing a model of the environment or multiple iterative policy evaluations.

3.7.4 Q-Learning

Iteratively updating and refining policies is foundational in RL, leading to a class of algorithms known as *fitted Q-value iteration*, with Q-learning being one such example. This algorithm marks a significant conceptual departure in its evaluation process: the new policy's value, π', is ascertained at the *subsequent* state s_i', rendering the learning process *off-policy*. This off-policy nature allows Q-learning to learn from data generated by a different policy, aiding in convergence and efficiency.

Earlier discussions on dynamic programming emphasised discrete state-action spaces. However, the *curse of dimensionality* [74, 75] reminds us that in most real-world scenarios, the state-action spaces become prohibitively large, making it computationally infeasible to represent value functions in a simple *table*. This challenge is well suited to ML methods, particularly deep learning, for effective function approximation. By integrating deep NNs with value iteration methods, we transition to a class of algorithms: *fitted value iteration*. An important example is deep Q-learning (DQN) [76], for which we provide an algorithm in Appendix B.3.

3.8 Policy Methods

While value-based methods focus on evaluating the worth or utility of different states or actions, there exists a more direct approach to RL. This approach seeks to directly learn the optimal policy. Consider a robot tasked with navigating from point A to point B in a grid city, as depicted in Figure 3.9. At each junction, the robot can either proceed straight, turn left, turn right, or move backward. The primary objective of

Algorithm 4: Q-Learning Algorithm for Optimal Policy Discovery: This algorithm iteratively approximates the optimal action-value function Q^* using a value iteration approach. Through successive updates, it converges to the optimal policy π^* that maximises expected rewards over time in a given environment. The process involves collecting experiences under a policy π, updating Q values based on observed transitions and rewards, and ultimately deriving the optimal policy from the learnt Q values.

```
Input: Environment with states $s$, actions $a$, and rewards $r$, discount factor $\gamma$,
       and a policy $\pi$
Output: Optimised policy $\pi^*$
Initialise $Q_\phi(s, a)$ arbitrarily for all state-action pairs
Collect dataset $\{s_i, a_i, s'_i, r_i\}$ by executing policy $\pi$ in the environment
for a predefined number of steps $K$ or until convergence do
    Set $y_i \leftarrow r(s_i, a_i) + \gamma \max_{a'} Q_\phi(s'_i, a')$ for each transition $(s_i, a_i, s'_i, r_i)$
    Update $\phi$ by minimising the loss: $\phi \leftarrow \arg\min_\phi \frac{1}{N} \sum_i ||Q_\phi(s_i, a_i) - y_i||^2$,
      where $N$ is the number of transitions in the dataset
Define $\pi^*(s) = \arg\max_a Q_\phi(s, a)$ for each state $s$
Return the optimal policy $\pi^*$
```

this agent is to minimise the total distance travelled, effectively reducing the number of decisions or actions made.

Framing this scenario in the context of a MDP, the states are represented by the grid positions, the potential actions are the directions, and a negative penalty is assigned as the reward for each step taken. A natural strategy to tackle such problems is to directly learn a policy that determines the robot's movement across the grid in a manner that ensures reaching the destination in the shortest time.

Mathematically, a rudimentary deterministic policy π might be articulated in terms of the advantage function A^π as:

$$\pi'(a_t|s_t) = \begin{cases} 1 & \text{if } a_t = \arg\max_{a_t} A^\pi(s_t, a_t) \\ 0 & \text{otherwise} \end{cases}$$

Historically, classical RL treatments rooted their discussions in such policy formulations, inspired by the foundational concepts of optimal control theory. This direct emphasis on determining policies without delving into their underlying value functions gave rise to **policy-based methods**.

In contrast to value-based methods (detailed in Section 3.7), policy-based methods seek to directly optimise the policy without the necessity of a value function. An intuitive policy might always instruct the agent to move in a direction that brings it closer to its end goal. This strategy doesn't necessarily require the calculation of cumulative reward over time. Instead, the policy is parameterised by certain parameters θ, and it is fine-tuned by adjusting these parameters in a manner that optimises a performance metric, $J(\theta)$. This adjustment process is an optimisation problem and is frequently addressed using gradient descent techniques.

The *policy gradient theorem*, fundamental to RL, provides a mathematical foundation for direct policy optimisation. In essence, it allows us to understand how small changes

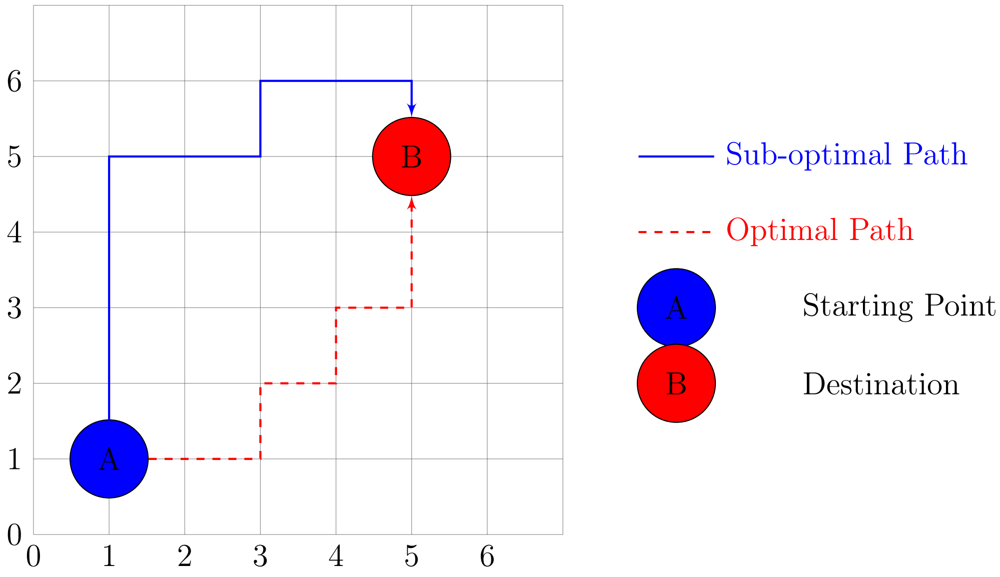


Figure 3.9: Illustration of a robot navigating a grid city. The robot starts at point A and aims to reach point B. The red path represents an optimal route (path length = 8) taken by the robot, while the solid blue path illustrates a possible sub-optimal route (path length = 10 > 8). The robot's objective is to minimise the distance travelled, choosing paths that are most efficient.

in policy parameters affect the expected return.

Theorem 3.8.1 (Policy Gradient Theorem)**.** *Within any MDP, applicable to both average-reward and start-state scenarios, the gradient of the performance function $J(\theta)$ with respect to policy parameters θ is proportional to the sum of state-action value functions weighted by the policy gradient. Formally, this relationship is expressed as:*

$$\nabla_\theta J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\boldsymbol{\theta}) = \mathbb{E}\left[\sum_a q_\pi(s,a)\nabla\pi(a|s,\boldsymbol{\theta})\right],$$

where the proportionality constant equals 1 for the continuous case and is equivalent to the average duration of an episode in episodic cases. In this context, $J(\theta)$ signifies the performance metric parameterised by θ, and $\mu(s)$ denotes the on-policy state distribution, indicating the frequency of occupying each state under policy π [64].

For a comprehensive mathematical exposition and derivation of the principles underlying this theorem, refer to Appendix B.4. This section provides detailed insights into the mathematical mechanics of policy optimisation in RL, facilitating a deeper understanding of the theorem's application and significance in various MDP contexts.

3.8.1 REINFORCE: Building on the Policy Gradient Theorem

Given the policy gradient theorem 3.8.1, one can derive a foundational classical algorithm - REINFORCE (a.k.a. Monte-Carlo Policy-Gradient Control). This algorithm is an intuitive extension of the principles set forth by the policy gradient theorem, primarily emphasising direct policy differentiation.

In this approach, the gradient of the performance measure, $J(\theta)$, is estimated using sampled trajectories. REINFORCE takes this a step further, using a Monte Carlo method to sample trajectories under the current policy π_θ and then estimate the gradient for policy optimisation.

A brief overview of REINFORCE is as follows:

1. Execute the current policy π_θ within the environment, producing a series of trajectories $\{\tau^i\}$.
2. Compute the gradient of the performance measure $J(\theta)$ with the help of these trajectories.
3. Adjust the policy parameters θ according to the estimated gradient to improve policy performance.

Referring to Figure 3.9, REINFORCE would iteratively hone the policy, transitioning from less effective paths to the most efficient trajectory by updating the policy based on reduced path lengths.

Algorithm 5: REINFORCE Algorithm for Policy Optimisation: This algorithm iteratively optimises the policy parameters $\boldsymbol{\theta}$ of a differentiable policy π using gradient ascent, based on the returns of sampled episodes and the likelihood of the actions taken. It effectively utilises the policy gradient theorem to find the direction that maximises expected rewards.

```
Input: Differentiable policy pasteurisation $\pi(a|s, \boldsymbol{\theta})$, Step size $\alpha > 0$, Discount
       factor $\gamma$
Output: Optimised policy parameters $\boldsymbol{\theta}$
Initialise policy parameters $\boldsymbol{\theta} \in \mathbb{R}^{d'}$
for each episode do
    Generate an episode $S_0, A_0, R_1, ..., S_{T-1}, A_{T-1}, R_T$ by following policy
      $\pi(\cdot|\cdot, \boldsymbol{\theta})$
    for $t = 0$ to $T - 1$ do
        Compute return $G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$
        Update $\boldsymbol{\theta}$ using the gradient ascent: $\boldsymbol{\theta} = \boldsymbol{\theta} + \alpha\gamma^t G_t \nabla_\theta \log \pi(A_t|S_t, \boldsymbol{\theta})$
Return the optimised policy parameters $\boldsymbol{\theta}$
```

The elegance of REINFORCE lies in its simplicity and straightforwardness among policy gradient methods. While it avoids the need for complex approximations, its reliance on the Monte Carlo method can result in inefficiencies in sampling, particularly in extensive environments.

3.9 Models in RL

Up to this point we have discussed value-based and policy gradient methods, mostly in the context of planning. The domain has been planning where the model is known, meaning the previous methods do not attempt to estimate the transition probability distribution or the reward function of the MDP. What happens when there isn't an explicit environmental model provided? How does one then efficiently navigate large state spaces? One could employ model-free techniques to directly learn a value function or policy, but this might reintroduce issues related to sample efficiency. An alternative is to first *learn* an environmental model and *only then* employ established planning tools, like value-iteration.

Returning to biological motivations, many intelligent agents, including humans, *learn* an explicit models of their environment — a representation or blueprint of how particular facets of the world operate [77, 78]. While the prior sections were entrenched in the learning of value functions or policies, typically facilitated by universal approximators, model-based RL introduces an augmentation: a *model* representing state dynamics. A shift in perspective occurs as we transition from the model-free paradigm, focusing on the immediate learning of policy or value function, to a model-based setting. Here, the priority is not just acquiring an accurate policy, but harnessing and refining a model that can inform planning.

This connects with the discussion on causality in the previous chapter. In causal modelling, it's important to clearly define and organise the assumptions that guide causal predictions. This method helps agents to methodically analyse how different environmental factors interact. A key challenge is accurately representing the relationships between these factors in the model. An advantage of causal models is that they can be tested and potentially disproven through empirical methods. This is what a causal agent does when it performs an intervention in its environment. Combining the tasks of (1) developing an accurate model, (2) making the best decisions over time, (3) learning from past actions, and (4) exploring sufficiently, requires careful thought.

Model learning and exploration, though intertwined, are distinct. An agent may possess qualitative knowledge about environmental dynamics without precise quantitative understanding regarding the associated uncertainties. The Taxi example from Section 3.5.1 illustrates this point. A taxi may possess a comprehensive environmental model but remain ignorant about the precise probability of finding a passenger at a given location. These nuances influence the planning phase — while the agent can still devise a plan, achieving optimal planning might remain elusive.

Classic model-based RL approaches can be categorised based on the nature of environmental dynamics:

1. **Deterministic:** In deterministic settings, the agent has full knowledge of environmental dynamics. Thus, given a particular state, the agent can confidently plan, anticipating environment's reactions to a series of actions. Formally, this can be represented as:

$$a_1, \ldots, a_T = \arg\max_{a_1, \ldots, a_T} \sum_{t=1}^{T} r(s_t, a_t) \text{ s.t. } a_{t+1} = f(s_t, a_t).$$

2. **Stochastic Open-Loop:** Here, the environment is inherently stochastic. An agent must remain committed to its plan from inception to completion, regardless of whether anticipated optimal conditions materialise. This can be mathematically described as:

$$a_1, \ldots, a_T = \arg\max_{a_1, \ldots, a_T} \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t)\right].$$

$$\begin{tikzcd}\text{Model} \arrow[r] & \text{Simulated Experience} \arrow[r, "\text{Backups}"] & \text{Values} \arrow[r] & \text{Policy}\end{tikzcd}$$

Figure 3.10: A diagram illustrating the flow from model to policy through simulated experience and values.

3. **Stochastic Closed-Loop:** This setting offers the agent the flexibility to adapt its plan based on observed dynamics. The primary objective then becomes:

$$\pi = \arg\max_{\pi} \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right].$$

3.9.1 Model Planning

Planning, especially within model-based RL, encapsulates the procedure of employing a known model to generate an improved policy. As the model comprehensively captures state dynamics, policy improvement converges to the identification of actions propelling an agent along the most reward-rich path in the state-space. The concept resonates with the *state-space planning* framework.

Distinct strategies for planning come into play when considering the optimisation of a sequence of actions, $A = a_1, \ldots, a_T$. A basic derivative-free strategy might involve randomly selecting actions and determining the optimal one. The significance of "derivative-free" methods lies in their capacity to handle non-differentiable, noisy, or discontinuous objective functions, offering a distinct advantage in certain RL scenarios. Numerous nuances and advancements built on this foundational idea will be discussed in the subsequent sections.

Rollout and Monte Carlo Tree Search

In decision-time planning algorithms, *rollout algorithms* are notable for their use in Monte Carlo control for simulating trajectories. Their main goal is to refine the accuracy of action-value estimates for a state by averaging returns from different action paths. These algorithms focus on identifying the best action for a current state using a given rollout policy, rather than estimating the optimal policy. Their effectiveness relies on the accuracy of action-value estimates and the depth of simulated future actions.

Monte Carlo Tree Search (MCTS) builds upon rollout algorithms by adding cumulative value estimation, leading to an approach that favours actions and simulations expected to bring higher rewards. MCTS has enhanced many state-of-the-art algorithms and operates through four main stages:

1. **Selection:** Starting from the root node, a tree policy guides the selection of a leaf node for expansion.

2. **Expansion:** Expansion reveals new actions at specific nodes.

3. **Simulation:** A rollout policy is used at the chosen leaf node to identify promising actions.

4. **Backup:** New action-value estimates are updated up the tree, ignoring values beyond the tree policy's scope.

This process repeats within computational limits, as illustrated in Figure 3.11. When it's time to select an action, a policy determines the choice. The sub-tree rooted at the current state can be reused from previous MCTS computations.

Algorithm 6: Generic MCTS Algorithm.

```
**Input:** Initial state $s_1$, Number of steps $K$
**Output:** Best action from $s_1$
for $i = 1$ to $K$ do
    Find leaf $s_l$ using TreePolicy($s_1$);
    Evaluate leaf with RolloutPolicy($s_l$);
    Update values between $s_1$ and $s_l$ in the tree;
Select the best action from $s_1$.
```

The tree policy for selecting leaf nodes is crucial for optimising performance. A common score function for choosing child nodes in the tree is:

$$\text{Score}(s_t) = \frac{Q(s_t)}{N(s_t)} + 2C\sqrt{\frac{2\ln N(s_{t-1})}{N(s_t)}},$$

where $N(s_t)$ represents the visitation frequency for node s_t, inspired by the UCB (Upper Confidence Bound) technique [79].

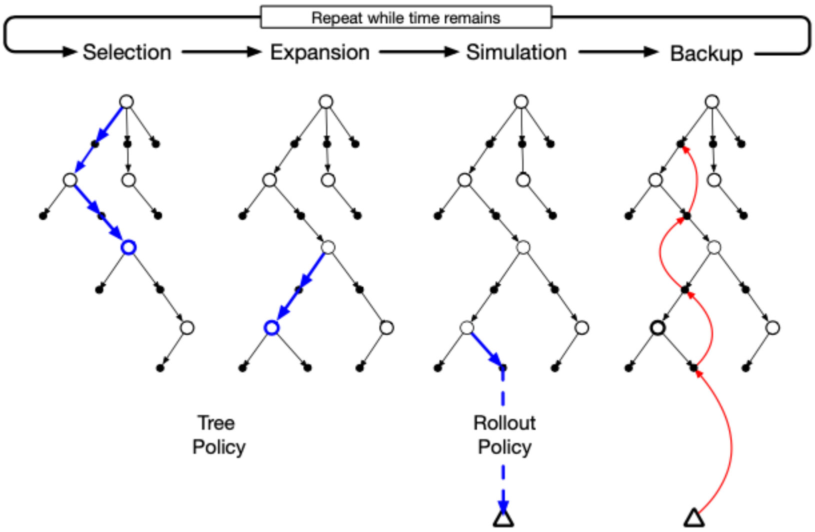


Figure 3.11: The stages of MCTS, as described by Sutton and Barto [64].

3.9.2 Planning in Model-Based RL

As already mentioned, an alternative to the model-free approach is to first *learn* an environmental model and *only then* employ established planning tools. Such a foundational approach to model-based planning is as follows:

This strategy has its roots in classical robotics for system identification, as illustrated by [80]. Nonetheless, when integrating deep learning techniques, its performance can degrade. This decline often arises from how data is collected according to a specific policy and then used to train the model. In situations where extrapolation from collected

Algorithm 7: Fundamental Model-Based Algorithm.

```
Input: Initial policy $\pi_0(a_t, s_t)$
Output: Strategy using $f(s, a)$
Apply initial policy $\pi_0(a_t, s_t)$ to accumulate a dataset $D = \{(s, a, s')_i\}$;
Use supervised learning to determine $f(s, a)$ by minimising:
  $\sum_i ||f(s_i, a_i) - s'_i||^2$;
Strategise actions using $f(s, a)$;
```

data becomes necessary, agents can make misguided decisions, leading to unintended outcomes, often termed as “falling off the cliff” in RL literature. One solution is data augmentation to cover the entire state space:

Algorithm 8: Augmented Data Model-Based Algorithm.

```
Input: Initial policy $\pi_0(a_t, s_t)$, Number of steps $K$
Output: Augmented dataset $D$ and strategy using $f(s, a)$
Apply initial policy $\pi_0(a_t, s_t)$ to accumulate dataset $D = \{(s, a, s')_i\}$;
for $i = 1$ to $K$ do
    Use supervised learning to determine $f(s, a)$ by minimising:
      $\sum_i ||f(s_i, a_i) - s'_i||^2$;
    Strategise for action $a'$ using $f(s, a)$;
    Implement action $a'$ and augment $D$ with the resultant data $\{(s, a, s')_j\}$;
```

While this technique harmonises with deep learning approaches, it remains susceptible to agents' errors. Re-planning after each data addition emerges as an effective strategy to counteract this vulnerability:

Algorithm 9: Dynamic Re-planning in Model-Based RL.

```
Input: Base policy $\pi_0(a_t, s_t)$, Outer loop steps $N$, Inner loop steps $K$
Output: Augmented dataset $D$ and strategy using $f(s, a)$
Apply base policy $\pi_0(a_t, s_t)$ to gather dataset $D = \{(s, a, s')_i\}$;
for $i = 1$ to $N$ do
    Use supervised learning to identify $f(s, a)$ by minimising:
      $\sum_i ||f(s_i, a_i) - s'_i||^2$;
    for $j = 1$ to $K$ do
        Strategise for action $a'$ using $f(s, a)$;
        Implement action $a'$ and append the outcome data $\{(s, a, s')_j\}$ to $D$;
```

Yet, a noticeable performance gap persists when compared to model-free methods. This disparity often stems from the model's inherent imperfections and its inability to account for unknown factors. The ensuing model might contain misleading local optima, leading agents to consistently opt for actions deemed high-reward based on flawed assumptions.

We now point the reader to Appendix B.6 for a brief discussion about uncertainty estimation in RL.

3.9.3 Model Learning

In model-based RL, it's vital to know how the environment reacts to specific actions. This knowledge allows agents to simulate experiences and choose the best actions for specific states, leading to more informed decisions. MBRL involves two primary steps:

1. Learning about the environment, called *learning*.
2. Using this knowledge to refine decisions, known as *planning*.

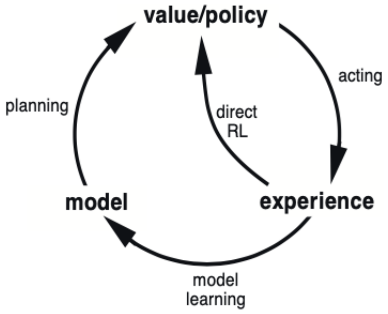


Figure 3.12: Model-based RL framework highlighting the relationship between actions, learning, and planning. More details in [64, Ch. 8].

Model-based methods excel at using limited data to form strategies. Yet, design limitations can sometimes make model-free methods more advantageous. Model-based RL is well-suited for enhanced decision-making because each data instance refines the model.

In the previous sections we considered planning using a learnt model. Those methods didn't need any policy to understand the environment. Instead, now we'll discuss an approach where we use the learnt model to develop a policy.

In the stochastic open-loop scenario, the agent doesn't realise it can modify future decisions, limiting its planning. As an agent learns more about its state at each time step, its expected rewards change. Earlier model-based RL strategies didn't account for this flexibility. The policy derived from such learning helps agents make decisions, either globally or by combining local policies.

To train policies using models, one might think of applying backpropagation to maximise our objective. However, this method can face issues, like vanishing or exploding gradients, especially in tasks that span long time horizons. An essential observation is that policy gradients don't rely on explicit gradient terms.

Integrating model-free RL techniques with model-based approaches leads to improved learning efficiency, particularly through the use of experience replay. Experience replay, a concept often employed in model-free RL, involves storing past experiences and then sampling from this pool to replay these experiences during the learning process. This strategy breaks the correlation in sequential experiences and allows for multiple learnings from individual experiences, thereby enhancing the learning efficiency.

When applied to model-based RL, experience replay can significantly enhance performance. This hybrid approach effectively combines the predictive power of a model

Algorithm 10: Backpropagation Model-Based Algorithm.

```
Input: Base policy $\pi_0(a_t, s_t)$, Outer loop steps $N$, Inner loop steps $K$
Output: Optimised policy $\pi_\theta(a_t|s_t)$ and augmented dataset $D$
Use base policy $\pi_0(a_t, s_t)$ to gather data $D = \{(s, a, s')_i\}$;
for $i = 1$ to $N$ do
    Minimise a loss, e.g., $\sum_i ||f(s_i, a_i) - s'_i||^2$ to find $f(s, a)$;
    for $j = 1$ to $K$ do
        Optimise policy $\pi_\theta(a_t|s_t)$ using backpropagation through $f(s, a)$;
        Execute $\pi_\theta(a_t|s_t)$ and add $(s, a, s')$ to $D$;
```

with the empirical robustness of model-free learning. In stochastic systems, both backpropagation and policy gradient methods, commonly used in model-free RL, have their respective advantages and limitations:

1. **Policy Gradient:** Offers more stability with an increased number of samples but tends to have high variance.
2. **Back-propagation:** While efficient, it can be susceptible to numerical issues over long-term problems. Early deviations can lead to significant discrepancies in later outcomes.

Dyna

This concept of enhancing model-based RL with model-free techniques, akin to experience replay, is exemplified in algorithms like Dyna, as introduced by Sutton [81]. Dyna combines model-based and model-free RL. It uses a learnt environmental model to simulate experiences (model-based aspect) and updates its policy based on actual experiences gathered from the environment (model-free aspect).

Algorithm 11: (Classic) Dyna.

```
Input: Current state $s$, Step size $\alpha$, Number of planning steps $K$
Output: Updated Q-values $Q(s, a)$
Pick action $a$ based on exploration (given state $s$);
Observe resultant state $s'$ and reward $r$ - $(s, a, s', r)$;
Update Q-value: $Q(s, a) \leftarrow Q(s, a) + \alpha\, \mathbb{E}_{s',r}[r + \max_{a'} Q(s', a') - Q(s, a)]$;
Refine models $\hat{p}(s'|s, a)$ and $\hat{r}(s, a)$ using new data $(s, a, s')$;
for $i = 1$ to $K$ do
    Draw $(s, a)$ from buffer;
    Update Q-values: $Q(s, a) \leftarrow Q(s, a) + \alpha\, \mathbb{E}_{s',r}[r + \max_{a'} Q(s', a') - Q(s, a)]$;
```

Dyna's structure is flexible. We can replace the model-free method or learn different models. Several modern algorithms, like MBA [82], MVE [83], and MBPO [84], use this Dyna-inspired structure. Interestingly, sometimes it's harder to train a model than a value function, making model-free methods a better choice in those cases.

3.10 World Models

World models have emerged as a popular method in RL. Unlike traditional model-based RL, which often presumes that agents can predict exact environmental dynamics, world models function as agents' internal representations, aspiring to understand the intricacies of their environments based on observational data. Ha and Schmidhuber [85] pioneered this approach, in which they emphasised the significance of agents building and utilising their internal representations of environments. Such representations, or world models, empower agents to simulate potential future trajectories, allowing them to envisage action sequences and their probable outcomes without immediate environmental interaction. The conceptual underpinnings of training using limited experience in a simulated environment are aligned with the current understanding of the role of sleep and dreaming in human cognition and skill acquisition [86, 87].

Conventional approaches in RL involve forming predictions to determine the action expected to provide the highest return, which is then selected, and the subsequent environmental response is used to refine the model. In [85] agents were trained to excel in the imagined world, with the ultimate aim being their effectiveness in the real world. In [88] the authors make several improvements to the world model formulation. The authors introduce an algorithm called **Dreamer** which has an identical interface for both the virtual and real environments, meaning policies mastered within the imagined realm could be seamlessly transferred to real-world scenarios.

Dreamer distinguishes itself from prior approaches by iteratively updating via backpropagation, a significant deviation from the latter's non-iterative approach. Dreamer's architecture can be summarised into three integral components:

1. **World model:** Utilises past experience data for training.
2. **Behaviour:** Employs the world model to derive behaviours purely from imagination, leveraging value and actor networks.
3. **Environment:** The agent's interactions with the environment augment the model's experiential dataset.

Dreamer uses an actor-critic approach that considers future rewards, not just immediate ones. It creates models of actions and states in a latent (hidden) space. The main task of the policy in this 'world model' is to choose actions that work best in a simulated setting. The goal is to have this policy work well in the real world. To achieve this, Dreamer trains both its action and value prediction models through a process of policy iteration.

As RL methods continue to evolve, especially those using world models, the link between causality and RL becomes an exciting area to explore. Counterfactual reasoning, key to understanding causality, aligns well with the concept of agents using their 'imagination', as seen in these methods. The ability for agents to think about 'what-ifs' – to imagine counterfactual outcomes – is crucial. It helps agents to not just move through but to understand and think deeply about their environments. Incorporating causal thinking could be a significant next step in making these internal world models of agents more detailed and reliable.

3.11 Multi-Agent Systems

This section extends the discussion from RL to environments with multiple agents. The recent rise of RL has brought renewed attention to multi-agent systems (MAS), leading to the development of both theoretical models and practical applications [89]. A key aspect of this resurgence is the shift towards decentralised learning in diverse settings.

3.11.1 Historical Context and Relevance

This section provides an overview of the historical development of MAS and their significance in contemporary technology, drawing from game theory, economics, and swarm intelligence.

Game theory, developed in the early 20th century [90], models strategic interactions among rational entities. Concepts like Nash Equilibrium, which describe scenarios where no participant can gain by solely changing their strategy, have been influential across disciplines like economics, political science, and biology. These ideas help explain a range of phenomena from market behaviour to evolutionary tactics [91].

Economic theories have also played a significant role in MAS, particularly in understanding markets like oligopolies and monopolistic competition. These models illuminate the dynamics of competitive and cooperative strategies in markets [92, 93].

Swarm intelligence, inspired by natural systems, studies group behaviours that emerge from individual interactions, such as ant trail formation or bird flocking. These principles, framed within MAS, are applied in various areas, including robotics and network systems [94].

The advancement of RL has brought new life to these classical models, adapting them to complex, dynamic contexts. MAS scenarios, unlike single-agent RL, encompass a range of interactions from cooperation to competition. The integration of RL into MAS, termed multi-agent RL (MARL), introduces novel opportunities and challenges. Game theory has been particularly important in MARL, aiding agents in navigating dynamic strategies [95–97].

In causality, Wright's path analysis in the 1920s laid the foundation for understanding causal relationships, later evolving into Bayesian networks in the 1980s. These concepts have been adapted to MAS, exploring the impact of agents' actions on each other in complex settings. Research in multi-agent causal models investigates cause and effect in environments with interactive and learning agents [22, 98].

Figure 3.13 presents a timeline illustrating significant milestones in these fields. It highlights critical developments like Nash's Game Theory and the progression of swarm intelligence, which have influenced multi-agent modelling. The timeline also shows the fusion of these classical models with contemporary RL, emphasising the shift towards dynamic, adaptive systems. This visual representation underscores the interdisciplinary nature of MAS research and informs ongoing and future studies in creating complex multi-agent environments.

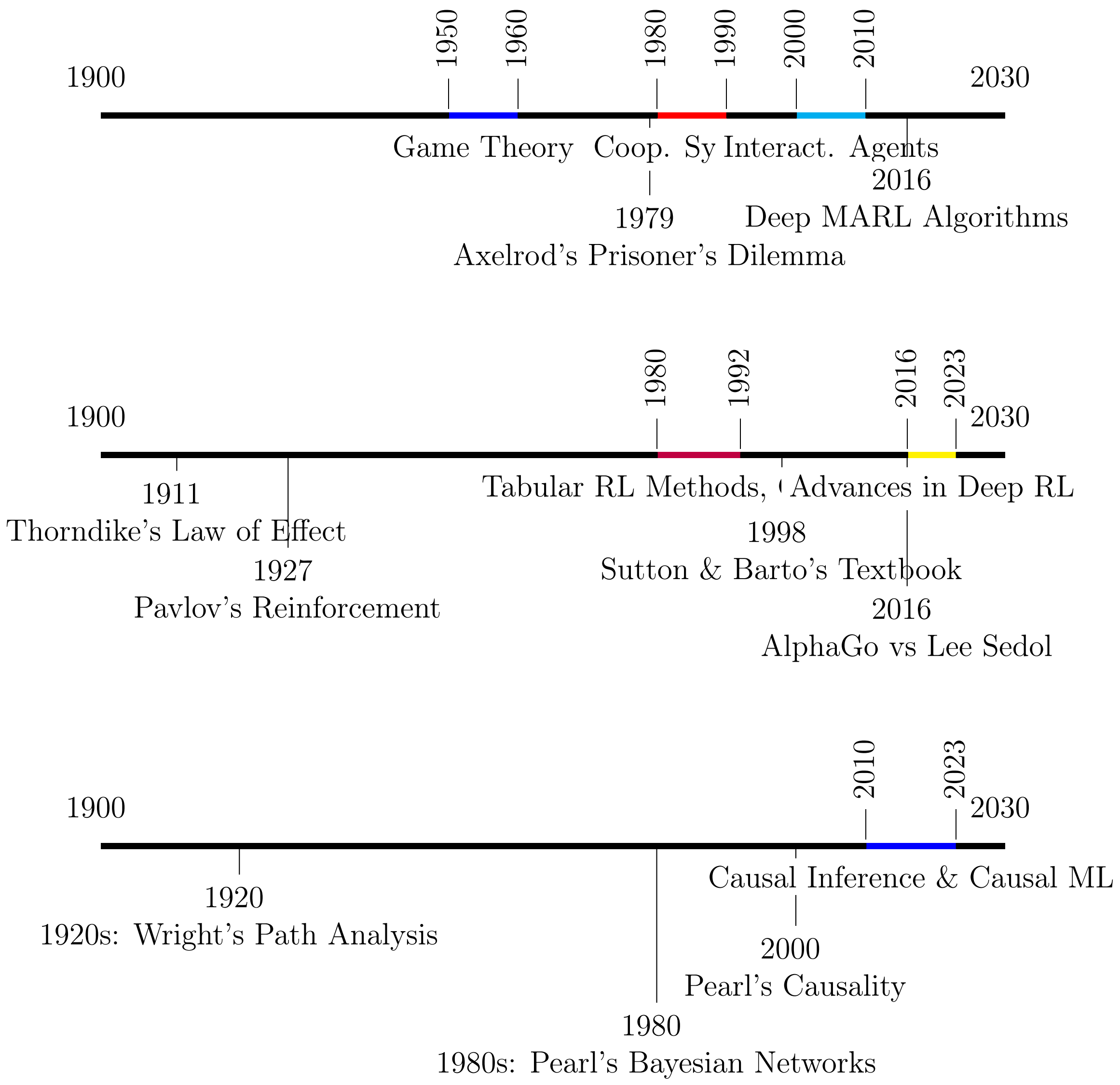


Figure 3.13: This figure presents three timelines illustrating key developments in game theory, RL, and causal inference from the early 20th century to the present. The first timeline (top) highlights milestones in game theory, beginning with the formalisation of Nash's game theory and evolving through the advent of interactive agents and deep multi-agent reinforcement learning (MARL) algorithms. The second timeline (middle) traces the evolution of RL, from early concepts like Thorndike's law of effect to recent breakthroughs exemplified by AlphaGo. The third timeline (bottom) charts significant advances in causal inference, from Wright's path analysis to modern applications in causal ML. These timelines depict a narrative of increasing interdisciplinary convergence, particularly in the last two decades, as techniques and theories from these fields begin to intersect and inform one another, suggesting a unified framework that bridges theoretical, algorithmic, and practical aspects across these domains.

3.11.2 Models in Multi-Agent Reinforcement Learning

Contemporary MARL methodologies primarily extend the MDP framework to accommodate the nuances of multiple agents interacting simultaneously. A seminal model in this domain is the Markov (or Stochastic) Game, introduced to the RL domain by Littman [95]. This model portrays systems transitioning in discrete steps, influenced by the collaborative actions of agents.

Definition 3.11.1 (Markov Game (or Stochastic Game))**.** *A Markov game encompasses a state space S and a collection of action sets $\{A_1, \ldots, A_k\}$ for each agent. State transitions are dependent on the current state and the joint actions of the agents, defined as $T = T(S, \{A_1, \ldots, A_k\}) \to PD(S)$. Here, $PD(\cdot)$ denotes a probability distribution over the state space. For each agent, denoted by i, there is an associated reward function:*

$$R_i : S \times A_1 \times \cdots \times A_k \to \{R_1, \ldots, R_k\}.$$

The inherent complexities of MARL arise not only from the dynamic interactions between the agents but also from variances in agent capabilities, access to information, and even individual objectives [99]. These complexities are further exacerbated in real-world scenarios that often present mixed interactions, comprising both cooperative and competitive elements [96, 97, 100–102]. Moreover, the curse of dimensionality presents significant challenges as the joint action and state spaces grow exponentially with the number of agents.

One fundamental challenge is the non-stationarity of the environment from any agent's perspective [103]. As agents continually adapt and learn new strategies, the environment, as observed by any particular agent, keeps changing, introducing added layers of complexity.

- **Cooperative Agents:** These agents work together to achieve a common goal or maximise a shared reward. The need for coordination, communication, and sometimes even prioritising the collective good over individual rewards becomes important. Inter-agent communication, especially in cooperative settings, is important. Developing effective communication protocols or allowing agents to evolve their communication strategies is often central to successful MARL applications.
- **Competitive Agents:** In contrast, these agents operate to outdo each other, aiming to maximise their individual rewards, often at the expense of others.
- **Mixed Scenarios:** Real-world applications often involve a combination of both cooperative and competitive elements. In such scenarios, agents must balance between collaborating with some agents while competing with others. For example, in RoboCup [104] where teams of agents compete in football.

Example 7 (Cooperative Scenario - Coordination Game)**.** *Consider two agents tasked with painting a large room. Both agents have a choice: to paint the left wall or the right wall. If both agents choose the same wall, the room gets painted twice as fast, resulting in a higher shared reward for both agents. However, if they choose different walls, the room still gets painted, but at a normal pace, leading to a lower collective reward. The optimal strategy here is for both agents to coordinate and paint the same wall together, thereby maximising their shared reward.*

Example 8 (Competitive Scenario - Prisoner's Dilemma)**.** *Two criminals are arrested, but the police don't possess enough evidence for a conviction. The police give each prisoner a choice: betray the other or remain silent. Both prisoners are in separate* *rooms, and they cannot communicate. If Prisoner A and Prisoner B both stay silent, they both serve a short sentence. If A betrays B but B remains silent, A goes free while B gets a long sentence, and vice versa. If both betray each other, both get a moderate sentence. Here, the strategy that maximises individual benefit is to betray, making it a competitive scenario.*

Safety and fairness are also of principal importance in MARL. Ensuring that agents operate without causing harm and that rewards or opportunities are equitably distributed across agents are emerging as vital research areas.

In many multi-agent environments, agents encounter partial observability, where they only have access to a subset of the overall state, known as an observation [105]. The Dec-POMDP model addresses this by focusing on cooperative scenarios where agents share both rewards and histories [106].

Definition 3.11.2 (Dec-POMDP)**.** *A Dec-POMDP is a 7-tuple:* $\langle S, \{A_i\}, T, R, \{\Omega_i\}, O, \gamma\rangle$ *for multi-agent contexts, where:*

- S *is the state space.*
- $\{A_i\}$ *represents the joint action set.*
- T *denotes transition probabilities.*
- R *is the reward function.*
- $\{\Omega_i\}$ *signifies the joint observation set.*
- O *depicts the observation distribution.*
- γ *is the discount factor.*

Example 9 (Dec-POMDP: Painting Coordination Game)**.** *Two agents are tasked with painting different parts of a room. Their visibility varies which influences their decisions. When both agents paint simultaneously, the work progresses faster. However, painting with poor visibility can lead to mistakes, resulting in penalties.*

*The **Formulation** of this scenario is as follows:*

*The **States,*** S*, comprise different visibility combinations for the agents. In terms of **Actions,*** $\{A_i\}$*, each agent has the option to either 'Paint' or 'Wait'. The **Transition Probabilities,*** T*, are based on the assumption that there are random transitions between these states. The **Reward Function,*** R*, assigns rewards or penalties to the agents depending on their decisions under the various visibility states. Agents receive **Observations,*** $\{\Omega_i\}$*, that inform them about the visibility of their respective sections. The **Observation Distribution,*** O*, determines which observation each agent receives, based on the current state. Lastly, the **Discount Factor,*** γ*, is set close to 1, highlighting the significance of future outcomes in the decision-making process.*

Using the painting coordination game as a reference, one might be tempted to simplify the problem by considering the agents as a single entity. This single-agent approach certainly has its merits but also comes with its own set of challenges.

One of the main advantages of a Single-Agent Formulation is the provision of a unified perspective. By considering multiple agents as a singular entity, this approach offers a holistic view of the system, ensuring that strategies and responses are consistent throughout. Furthermore, this formulation simplifies coordination. The removal of the need for inter-agent communication means the system has the potential to make decisions in a swifter and more cohesive manner.

On the other hand, the challenges of a single-agent formulation are multifaceted. The foremost challenge is the explosion of state and action spaces. When you combine the states and actions of various agents, the result is a combinatorial increase in possibilities. This can make the problem exponentially larger and more intricate, especially as the number of agents or their potential states and actions increase. Another challenge is the loss of individual nuances. A model that leans too heavily on a singular representation might inadvertently overshadow the unique constraints, behaviours, or specialities of individual agents. This can lead to inefficiencies or even errors in the system's operations. Scalability is also a concern. As more agents are integrated into the system, the complexity of modelling them as a single entity grows exponentially, presenting significant computational challenges. Lastly, the reduced generalisability of this approach is evident when we consider diverse scenarios. A solution tailored for a combined entity might struggle to adapt in situations where agents have varied configurations or distinct roles.

The decision to opt for a multi-agent vs. single-agent formulation often boils down to the specific requirements of the problem. If the agents have significant independent roles or unique constraints, a multi-agent approach might be more suitable. However, for tasks where agents have similar roles and a unified strategy is preferred, a single-agent perspective might be sufficient. The computational complexity of the problem, especially with increasing agent numbers, is also a determining factor. While a single-agent model simplifies coordination, it might become computationally infeasible for larger systems. It is therefore essential to weigh the advantages against the challenges to determine the best approach for the task at hand. While single-agent representations might seem simpler at a glance, they often prove infeasible for larger or more diverse agent systems. On the other hand, Dec-POMDPs offer a more scalable and realistic modelling approach, particularly for systems with unique agent identities and partial observabilities.

At this point, we refer the reader to Appendix B.5 for a selective presentation of some popular MARL algorithms.

3.12 Conclusion and Next Steps

In this chapter, we discussed how difficult it is to define intelligence and how this makes it hard to *design* or intelligent agents. We talked about deep learning and its effectiveness due to universal approximation theorems. We also mentioned the difference between clear-cut causality and the more opaque nature of NNs, giving a multi-layered view on this topic.

Then, we focused on RL, explaining its basic concepts and exploring value-based and policy methods. In the context of model-based RL, we explored how agents can learn and plan using models to perform RL tasks. We introduced the idea of world models, linking back to our earlier discussions about causality.

The chapter ended with a look at multi-agent systems. We provided some historical background, talked about the challenges in modelling these systems, and gave an example of a multi-agent RL algorithm. As we move to the next chapter, we aim to blend the ideas from the chapters on causality and RL. A key part of this integration is understanding action and intervention. Our goal is to create agents that are not just intelligent but also robust, adaptable, and capable of complex reasoning. The next chapter will explain how we can do this by combining the concepts of causality and RL.

Chapter 4

Causal Reinforcement Learning

Although both causal inference and RL have individually received considerable attention, their intersection remains relatively unexplored [107]. This intersection, rich with potential, directly relates to **RQ1**: *Do existing RL methods exhibit causal understanding?* In this chapter, we consider this question, investigating how causal mechanisms can potentially enhance RL algorithms. Emerging research indicates that integrating causal mechanisms into RL frameworks can notably mitigate inherent limitations of traditional RL algorithms [63, 108–112]. This promise has been one focus of my prior research [3] which served as the motivation behind this MSc investigation.

Central to the both these domains is counterfactual reasoning. As we discussed in detail in Chapter 2, the formal study of counterfactual reasoning has been pioneered by Judea Pearl [15]. We argue that the resultant theory offers RL methods improved model robustness, explainability, and versatility, particularly under varying conditions such as offline and off-policy scenarios.

This chapter serves three primary objectives:

1. To elaborate on the methodological approach taken for this extended literature review.

2. To dissect the specific methodologies through which causal inference can be integrated into RL, enhancing both model adaptability and computational efficiency.

3. To highlight the reciprocal benefits that RL techniques offer to causal inference, particularly in the context of data efficiency and robust estimation methods.

4. To touch on works that, although not explicitly at the junction of RL and causality, offer valuable techniques or theoretical frameworks that are pertinent for advancing CRL.

Incorporating causal reasoning into multi-agent systems introduces an additional layer of complexity, raising compelling research questions particularly around collaborative modelling and decision-making. These issues will also be discussed as they are relevant for investigating **RQ1** and **RQ2**.

4.1 Motivation

In domains such as healthcare and finance, the Markov property's assumption — that the future state only depends on the current state and action — is often inadequate. This is highlighted in the Partially Observable Markov Decision Process (POMDP) framework, where decisions are based on both current and historical states.

The Dynnamic Treatment Regime (DTR) in healthcare serve as a prime example of this inadequacy [113, 114]. In DTRs, interventions must account for a patient's comprehensive medical history and unique characteristics, not just their current health state. We define a DTR more formally in Section 4.4.2. The presence of many unobserved latent variables, which can significantly impact treatment outcomes, further complicates this scenario [115, 116]. The connection between DTRs, causal models, and RL is explored, given these complexities.

In POMDPs, the transition to a future state s' depends on the current state s, action a, and history h: $P(s'|s, a, h)$. This interplay of observed and latent variables necessitates a more comprehensive approach than traditional Markovian models provide. Causal theory, with its focus on disentangling dependencies between variables, is particularly suited for this challenge [117, 118].

The limitations of the Markov property in fields like healthcare and finance underline the need for a more holistic approach, such as causal reasoning. This approach is essential for understanding the intricate interactions between variables across different histories. The following section delves into the intersection of causality and RL, building on this motivation.

4.2 Methodology: The Landscape of Causality and RL

This section outlines the methodology used to examine the intersection of causality and RL in academic literature. The goal was to understand their overlaps and impacts, leading to insights into current trends. The methodology adopted in this literature review is designed to shed light on the gaps and opportunities in the field of causal RL, directly contributing to our understanding of **RQ1** and **RQ2**. By focusing on literature that intersects causality and RL, we aim to uncover the extent to which current RL methods embody causal understanding and the potential for causal models to enhance RL in complex environments.

The research began by identifying key terms like "Causal Inference", "Reinforcement Learning", and "Bayesian networks". The phrase "causal reinforcement learning" was particularly fruitful, revealing significant work by Bareinboim's Causal AI lab at Columbia University. Additionally, multi-agent research was found to be heavily influenced by game theory. Searches were conducted in academic databases such as Google Scholar and arXiv using these terms.

The selection of literature focused mainly on articles published after 2016 to align with the advancements in deep learning. Priority was given to works that combined causal reasoning with RL. The high quality of the identified literature necessitated minimal exclusion. Data from these articles were extracted to identify prevalent research questions and methods, revealing existing gaps in the field.

The data were categorised by their focus on causality, RL, or their intersection. This categorisation led to a synthesis that highlighted important trends and challenges. Influential contributions from researchers like Judea Pearl and institutions such as the Montreal Institute for Learning Algorithms (MILA) were significant in shaping the review.

Key Trends The integration of causal inference, particularly graphical methods, with RL algorithms has improved performance in various sectors. Another trend is the use of counterfactual reasoning in RL, which has led to more efficient strategy assessments. Combining deep learning with model-based RL is fostering advanced data integration and understanding, with multi-agent RL in cooperative settings benefiting from causal insights. The identified trends and emerging research questions reflect the core objectives of this thesis, particularly resonating with the themes of **RQ1** and **RQ2**. By analysing how causal inference is being integrated into RL and observing its impact on multi-agent systems and complex environments, we contribute to the broader discourse on the potential of causal models in enhancing RL methodologies.

Research Questions Emerged. The review generated several important research questions:

1. How can RL agents identify causal paths in mazes with complex variables like weather conditions or other agents?
2. How does the combination of imitation learning and RL in multi-agent settings aid in knowledge sharing, specialisation, and collective exploration?
3. In multi-agent systems, how can developing partially complete but causally accurate models facilitate the merging of collective knowledge?
4. Can integrating intrinsic motivation with robust RL, supported by causal structure discovery, improve agents' exploration in novel areas?
5. Is it possible to create a learning mechanism where agents alternate between causal structure exploration and policy optimisation?
6. What are the effects of learning causal models through unsupervised methods on understanding their significance?
7. What impact do incorrect causal models have on RL performance, and what theoretical aspects are involved?
8. Can complex games like Bridge be used to evaluate causally accurate partial models in multi-agent research for causality and RL?

Exploring the relationship between causality and RL highlights the challenge of partial observability. In many real-world situations, agents lack complete environmental information. This issue, framed within the POMDP model, contradicts the assumptions of traditional MDPs. In this context, agents must consider their entire interaction history and deduce latent factors affecting their observations. This mirrors the goal of causal discovery, which seeks to understand the underlying structure of data generation. The next section on partial observability will delve into the connection between probabilistic correlations and causal interpretations, aiming to provide an in-depth view of the challenges and potential solutions in this area.

4.3 Partial Observability

Building upon the POMDP concept introduced in Section 3.5.1, the lack of complete observability in many real-world scenarios renders the Markov property inapplicable. In contrast to the simplified decision-making in standard MDP settings discussed in Section 3, agents in POMDP settings could be required to sift through their entire interaction history to make informed decisions, depending on the chosen approach.

As we have argued, the computational challenges posed by POMDPs necessitate an exploration of causal approaches to the learning and reasoning task. Specifically, POMDPs compel us to interrogate: *What latent factors might have contributed to the observed state?* This mirrors the underlying query of causal discovery: *How is the underlying data-generating process structured to yield the observed data?* [119].

Aligning with the insights from Gershman [120], there is an acknowledgement within the RL community that viewing POMDPs as probabilistic causal models could be theoretically useful. This paradigm shift goes beyond mere probabilistic correlations, opening avenues for nuanced causal interpretations [117]. In POMDPs, observations are not mere data points; they embody events moulded by latent causal variables. Various Bayesian techniques facilitate the inference of such latent factors, unveiling the underlying generative model [121]. Transitioning towards causal approaches enriches traditional model-free RL approaches by pivoting away from a purely behaviourist foundation towards a more cognitively-rich, causally-informed paradigm [122]. This exploration into the causal implications within POMDPs underlines the primary inquiry of our **RQ1**: *Do existing RL methods exhibit causal understanding?* By investigating the latent factors and their causal influences in POMDPs, we delve deeper into the capabilities and limitations of current RL methodologies in embodying causal reasoning.

Figure 4.1 illustrates the nuanced relationships between probabilistic transitions and causal influences in POMDP settings. The narrative of probabilistic and causal connections is further illustrated through a depiction of state dependencies over time in a non-Markovian setting in Figure 4.2.

The consideration of partial observability in complex environments, as considered in Liang and Boularias [123], aligns with our **RQ2**. It investigates whether a causal model can enhance the coordination and sample efficiency of learning agents, particularly in decentralised learning tasks where partial observability plays an important role.

Liang and Boularias [123] present a creative approach to tackling the challenges of partial observability. At first, their algorithm follows the Markov assumption, but as it progresses, it brings in hidden variables similar to memory units in RNNs [61]. These variables encode information about past events, helping to ease the issues caused by partial observability. Metrics of information gain highlight key events, while independence tests help identify causal relations between variables.

Foundational to all of this is *available data*. The availability of large datasets has

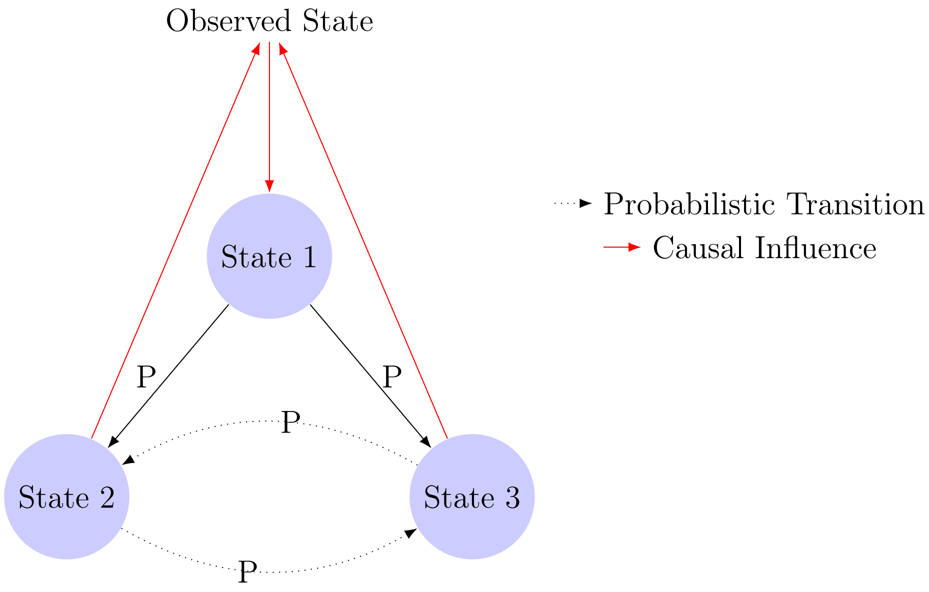


Figure 4.1: Visualisation of Probabilistic vs. Causal Relationships in a POMDP. In the figure, circles represent states in the environment, with State 1 being a focal state influenced by the Observed State. The arrows indicate relationships between these states. Dotted arrows depict probabilistic transitions, signified by the letter P, suggesting possible state transitions with associated probabilities. For instance, the transition from State 1 to both State 2 and State 3 is based on some probability. The dotted arrows between State 2 and State 3 indicate that transitions can also occur between these states, emphasising the intricate probabilistic interplay in POMDPs. In contrast, the solid red arrows show causal influences. The Observed State has a direct causal effect on State 1, and both State 2 and State 3 causally influence the Observed State. This distinction underlines the dual nature of relationships in POMDPs, where states can be connected through both probabilistic transitions and direct causal effects.

driven much of ML progress. That said, only recently has attention shifted to offline RL, where RL agents can learn from stored data. We now consider some of theses ideas, highlighting what makes the problem difficult, and one might resolve the issues that arise.

4.4 Generalised Policy Learning

Modern RL methods are often criticised for their heavy reliance on high-quality data and a preference for online learning [124]. Online policy learning, with its changing data and lack of clear indicators for these changes, requires agents to be flexible and resilient. While offline learning is discussed in the literature, it doesn't receive as much focus as online learning. Traditional RL agents, often working in isolation, use substantial computational resources. Offline RL, in contrast, aims to blend the proactive aspect of RL's learning with the proven success of classical statistical methods on fixed datasets.

At first glance, offline policy learning, which uses a fixed dataset, may seem to resemble model-based RL. Here, *static* means the data is not from ongoing interactions with the environment, but it may include interventional data. A key distinction lies in how the data is generated. Investigating the data generation process aligns with causal inference principles, suggesting that framing these issues with causal models could be beneficial.

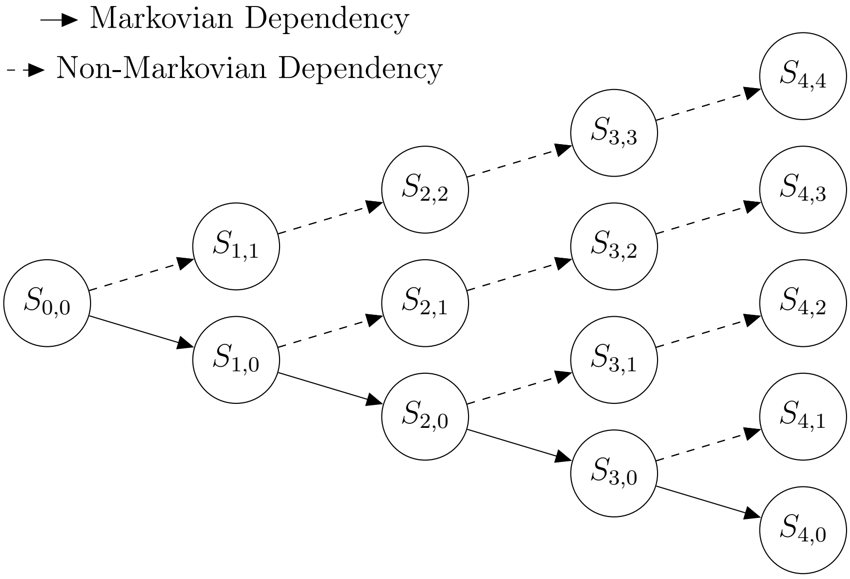


Figure 4.2: Visualisation of state dependencies over time in a non-Markovian setting. Circles represent states at different time steps, with solid arrows indicating the primary Markovian dependencies (where a state depends on its immediate predecessor). Dashed arrows depict non-Markovian dependencies, illustrating how each state can be influenced by multiple preceding states. As time progresses, the density of these non-Markovian dependencies increases, underscoring the complexity introduced when states are not Markovian.

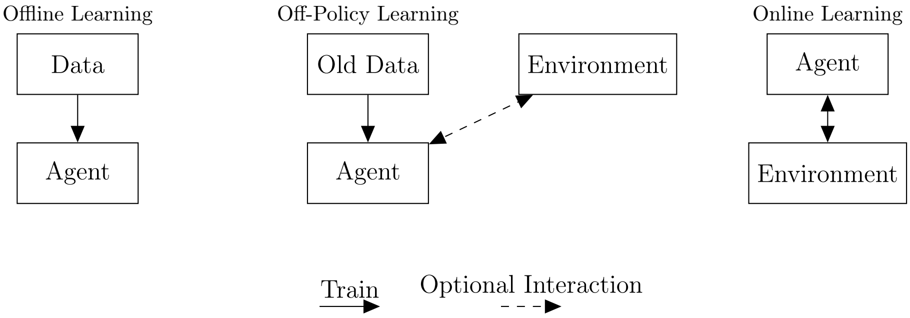


Figure 4.3: A graphic delineating the differences between offline, off-policy, and online learning modes. In offline learning, an agent is trained purely on existing data. In off-policy learning, the agent learns from old data but can optionally interact with the environment under a different policy. In online learning, the agent continually interacts with the environment, updating its policy based on real-time feedback.

4.4.1 Exploitation of Causal Inference and Transfer Learning

Conventional RL agents are subjected to training in isolated environments, utilising considerable computational and energy resources. As discussed, offline policy learning is centred around deriving insights from a static dataset, while online policy learning entails real-time learning, substantially constrained by time. Due to the evolving nature of data in online scenarios, agents are required to exhibit a degree of flexibility. The training phase for modern state-of-the-art agents can be extensively time-consuming.

Transfer learning emerges as a remedy to this learning inefficiency by utilising previous knowledge and experience to boost learning performance, similar to how humans apply past knowledge in confronting new tasks [125–127]. This aspect will be discussed in greater depth in Section 4.8. Causal inference also traverses a similar challenge of discerning effects from diverse data sources, with a significant impediment being learning amidst unobserved (hidden) confounders. In this context, 'causal bounds' refer to the bounds on the parameter space of the causal effect $E[Y|z]$ derived from the observational data $P(x, y|z)$. This concept is critical when precise estimation of target quantities is unfeasible, providing a range within which the true causal effect is likely to lie. In this subsection, we explore the potential of applying causal models for the Multi-Armed Bandits (MAB, see Appendix B.1) and MDPs problem setting, with the intention to enhance learning performance through a combination of observational and interventional learning modes.

An approach to tackling this challenge is discussed in Zhang and Bareinboim [128], where the authors combine transfer learning in elementary RL with causal inference theory. This integration is illustrated in a setting involving two MAB agents provided with a causal model of the environment. In scenarios where causal effects are elusive, the methodology proposed for deriving causal bounds from the knowledge encapsulated in available distributions is particularly pertinent. The use of methods of causal inference for aiding policy learning, as we venture into observational and interventional learning modes, directly pertains to our **RQ1** and **RQ2**. This approach differs from the traditional RL frameworks and seeks to answer if existing methods can be enhanced for better causal understanding and efficiency, as posited in our research questions.

The discourse then shifts towards contextual bandits, a variant of MABs (see Appendix B.1), discussed in [129], where the agent perceives additional contextual information correlated with the reward signal. Zhang and Bareinboim [128] commence with an off-policy learning scenario where agent A adheres to a policy denoted by $do(X = \pi(\epsilon, u))$, with context $u \in U$, outcome y, and noise ϵ, leading to a joint distribution $P(x, y, u)$. Another agent, denoted A', tries to learn about the environment and capitalise on A's experience to expedite its learning and promptly converge to the optimal policy. This tasks boils down to discerning the causal effect of an intervention on X, expressed as $\mathbb{E}[Y \mid do(x)]$.

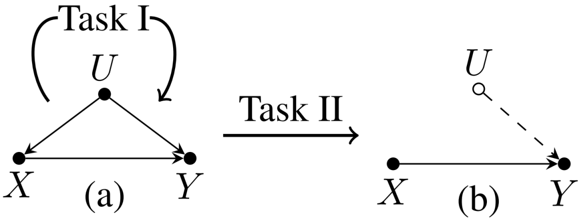


Figure 4.4: Figure extracted from [128] demonstrating the transfer learning task (II) between two MABs. (a) illustrates a causal graph with known context, U, while (b) portrays a standard MAB with unobserved confounder (unknown context) indicated with a dotted directed edge.

One might hastily conclude that if the identifiability condition is not satisfied, the preceding data is of no use in the transfer process. However, [128] reveals that even for non-identifiable tasks, it's possible to obtain causal bounds over the expected rewards for the target agent (Theorem C.1.1). At this point we refer the reader to the appendix for a theoretical look at regret bounds and a presentation of an accompanying algorithm (B-kl-UCB).

4.4.2 Tackling Dynamic Treatment Regimes

Zhang and Bareinboim [130] consider DTRs in the context of personalised medicine, exploring the potential of online RL algorithms in selecting optimal DTRs based on observational data. They hope to leverage the success of RL in enhancing decision-making processes in other fields for DTRs. Essentially, the aim is to find an optimal policy π that optimises a certain outcome Y, typically the patient's recovery or improvement in health markers, under the constraints of the DTR. However, the unknown parameters of the DTR often thwart direct optimisation attempts. Traditional algorithms necessitate the absence of unobserved confounders, and the common randomisation techniques usually don't sit well in the medical domain due to the risks involved. RL shines in this scenario by promising an efficient learning of DTRs while judiciously exploring the state-space and harnessing rewards. We provide a definition for a DTR below for reference.

The investigation into Dynamic Treatment Regimes (DTRs) and their optimisation through RL algorithms speaks directly to **RQ2**. By examining how causal models can potentially improve learning efficacy in these multi-staged decision-making processes, we contribute to understanding the broader implications and applications of causal RL in real-world scenarios.

Definition 4.4.1 (Dynamic Treatment Regime [130])**.** *A Dynamic Treatment Regime (DTR) can be formalised as a SCM defined by $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{u}) \rangle$, where $\boldsymbol{V} = \{\overline{\boldsymbol{X}}_K, \overline{\boldsymbol{S}}_K, Y\}$ encapsulates the total stages of interventions over the course of the regime. In this setup, $\overline{\boldsymbol{X}}_K$ signifies the progression $\{X_1, \ldots, X_K\}$ across stages. For each stage $k = 1, \ldots, K$:*

1. *The decision variable X_k is determined by a policy function, expressed as $x_k \leftarrow f_k(\overline{\boldsymbol{s}}_k, \overline{\boldsymbol{x}}_{k-1}, \boldsymbol{u})$, where X_k is finite and directed by a behavioural strategy.*
2. *The state variable S_k is derived through a transition function, denoted as $s_k \leftarrow \tau_k(\overline{\boldsymbol{x}}_{k-1}, \overline{\boldsymbol{s}}_{k-1}, \boldsymbol{u})$, where S_k is a finite state.*
3. *The final outcome Y at stage K is obtained via a reward function, formalised as $y \leftarrow r(\overline{\boldsymbol{x}}_K, \overline{\boldsymbol{s}}_K, \boldsymbol{u})$, and is constrained within the range $[0, 1]$.*

The exogenous variable values in $\boldsymbol{U}$ are sourced from the distribution $P(\boldsymbol{u})$.

Despite its popularity, the typical RL techniques fall short in the DTR context because of their dependency on the Markov property, which DTRs don't adhere to. The treatment at any given stage in DTRs is influenced by past treatments, contradicting the Markovian assumption of memorylessness. The authors reformulate a DTR as a SCM encompassing several stages of interventions. Each stage in a DTR consists of a decision, a state transition, and ultimately, an outcome Y at the final stage, determined by a reward function.

A specified DTR M^* results in a particular observational distribution which governs the data observed without any intervention. A policy π within the DTR defines a sequence of interventions that lead to an interventional distribution. The expected cumulative reward, denoted by $V_\pi(M^*)$, hinges on the policy π and the ultimate goal is to uncover the optimal policy π^* that maximises this expected reward.

To tackle the optimisation of an unknown DTR, the authors introduce the UC-DTR algorithm. This algorithm embodies an "optimism in the face of uncertainty" stance, a popular approach in the RL literature. With only the knowledge of state and action domains at its disposal, UC-DTR impressively manages to achieve near-optimal total regret bounds swiftly. The algorithm operates by proposing a new policy π_t based on collected samples up to the current episode t. Using the empirical estimates for expected reward and transition probabilities, it contemplates a range of plausible DTRs and computes the optimal policy for the most optimistic DTR in this set. This process iterates until a predefined tolerance level or episode count is reached.

Beyond the challenges introduced by more complex models and policies, there exist challenges in the well studied MDP domain. Selecting where and when to intervene – by taking an action – is a tough theoretical and empirical challenge. We now investigate this.

4.5 Selecting Points of Intervention

A recurrent issue in the domain of RL literature deals with managing the delicate balance between exploring the state-action space to evaluate long-term interventional outcomes, and exploiting current knowledge about the system's behaviour optimally. Recent studies have gravitated towards understanding the impact of non-trivial dependencies among the bandit's arms, which is now known as *structural bandits*. Herein, the causal relationships among these dependencies are portrayed using causal graph structures. For instance, the work by [131] discuss RL within a framework of causal models with unobserved confounders, a topic further explored in section 4.6 below. The crux of the argument is that counterfactual quantities can be instrumental in accounting for unseen confounders, enabling a robust policy convergence where traditional approaches falter.

This section is dedicated to identifying the most strategically beneficial action within a MAB framework, by correlating the selection of a bandit's arm with an intervention in a specific causal graph. This approach utilises the understanding of causal systems to identify what we term **optimal interventional choices**. The exploration of structural bandits within the SCM-MAB framework ties directly into our **RQ1**, probing into whether existing RL methods, particularly in the context of intervention selection, exhibit a comprehensive causal understanding. To elaborate on the intersection of MABs and causal inference, we introduce the notion of SCM-MAB as defined below:

Definition 4.5.1 (SCM - Multi-Armed Bandit (SCM-MAB) [132])**.** *Consider an SCM, represented as* $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U})\rangle$*, where* Y*, belonging to* $\boldsymbol{V}$*, serves as the reward variable within the domain* $\mathbb{R}$*. The arms of the bandit are conceptualised as* $\{\boldsymbol{x} \in Dom(\boldsymbol{X}) \mid \boldsymbol{X} \subseteq \boldsymbol{V} \setminus \{Y\}\}$*, signifying possible interventions on the SCM's endogenous variables excluding the reward. Associated with each arm is the reward's interventional distribution,* $P(Y \mid do(x))$*, and its mean, defined as* $\mu_x = \mathbb{E}[Y \mid do(x)]$*.*

While the SCM-MAB model gains insights from the causal graph and the associated rewards, it does not have knowledge of the SCM's functional mappings, $\boldsymbol{F}$, nor the joint distribution of the exogenous variables, $P(\boldsymbol{U})$. The objective, therefore, is to determine a minimal set of interventions that optimise action selection. This issue has been discussed in the referenced literature, though it falls outside the primary focus of this thesis. For more comprehensive definitions and further understanding, the reader is directed to the appendix.

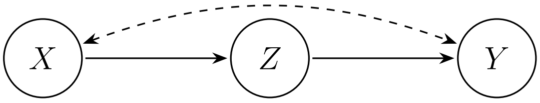


Figure 4.5: Illustration of a basic causal model with an unobserved confounder influencing X and Y. Here, intervening on Z emerges as potentially optimal. However, if Z intervention is infeasible, X intervention, alongside non-intervention, might be deemed possibly optimal, thus expanding the POMIS definition (C.2.2, see appendix for details).

Pinpointing the 'when' to intervene remains a challenge. Assuming an agent accrues causal knowledge progressively, determining when to switch from exploration to exploitation, or to intervene, embodies a crucial question. The literature illustrates mechanisms like confidence bounds to ensure a methodological transition from exploration to exploitation, an aspect that harmonises with causal reasoning. With ever-evolving causal knowledge, an agent ought to adjust its actions to mirror newfound insights. In this sense, both the 'where' and the 'when' of intervention require thought to ensure an agent's optimal performance. This dynamic decision-making process resonates with our **RQ2**, investigating the potential of a causal model to enhance the coordination and efficiency of learning agents, especially in the context of optimally selecting intervention points.

4.6 Counterfactual Decision Making

In traditional RL, decisions are made based on observed experiences and the expected future rewards of those experiences. However, real-world scenarios often present complexities that extend beyond the scope of these models. What if an agent could reflect upon its actions, considering alternative paths and their potential outcomes? Such hypothetical scenarios, where an agent ponders over the "what-ifs" of its decisions, are counterfactual problems.

This process of counterfactual thinking is not alien to human cognition. People often engage in counterfactual reasoning, contemplating scenarios like

> *Had I chosen a different major in college, where would I be now?*"

or

> *If I had invested in that stock five years ago, how wealthy would I be?*"

Such introspections can lead to better decision-making in future scenarios, learning from past experiences, and understanding missed opportunities. Counterfactual thinking in humans is integral to learning and problem-solving, enabling individuals to understand causality and consequences of actions, ultimately influencing future decisions [133, 134].

Furthermore, the concept of dreaming in humans can be seen as a sophisticated form of counterfactual thinking and the development of world models. Dreams often involve scenarios that are not part of our waking experiences, allowing the mind to explore outcomes and situations beyond the constraints of reality. This process can be viewed as an advanced form of counterfactual reasoning, where the brain simulates various scenarios, possibly as a means to prepare for future challenges or to process past experiences. This aligns with the theory that dreaming plays a role in emotional regulation and problem-solving [135, 136].

In a similar vein, one could posit that RL systems could employ counterfactual reasoning to construct and refine internal models of their environment, akin to human dreaming. By simulating various scenarios and outcomes, these systems can enhance their decision-making capabilities, especially in stochastic and complex environments. This approach could lead to more adaptable and resilient AI systems, capable of navigating diverse and changing situations with greater efficacy. Given this motivation, we now consider the role that causal inference could play in explicit counterfactual decision making.

4.6.1 The Role of Causal Inference in Counterfactual Decision Making

Causal inference stands out for its proficiency in addressing counterfactual queries. In RL, where learning is approached through interventions in a trial-and-error fashion, the integration of causal inference methodologies presents an ideal framework for tackling problems in counterfactual decision-making. This is particularly relevant when unobserved confounders influence decision-making processes. This integration idea brings us back to **RQ1**: *Do existing RL methods exhibit causal understanding?* The ability of RL agents to engage in counterfactual reasoning showcases a significant step towards causal understanding within the RL framework.

Within RL, the ability to make counterfactual decisions potentially elevates an agent's learning capacity, enabling it to derive insights not only from its actual experiences but also from potential actions it has not considered. In theory, this capability empowers agents to predict the results of unexecuted actions and modify their strategies accordingly, based on these envisioned scenarios. This type of decision-making is of paramount importance in environments where choices have long-standing impacts and where testing every possible action is not practical.

4.6.2 Augmented Bandit and RL Models

To enhance RL with counterfactual reasoning and causal inference, it is necessary to modify and expand our existing models. An essential modification is the inclusion of models that consider unobserved confounders—variables that impact both decision-making and outcomes, yet remain unobserved. We start by examining a simple bandit problem, and then extend our insights to MDPs.

A key intersection between RL and causal inference is the objective of regret minimisation. In the context of MABs, policy strategies are formulated to reduce regret over time. The principle is illustrated in the study by Auer et al. [137], who introduced the *UCB2* policy. Their research shows that, under specific conditions, logarithmic regret can consistently be achieved over time.

The mathematical bound for expected regret after numerous plays is described as

$$\sum_{i:\mu_i<\mu^*} \left(\frac{(1+\alpha)(1+4\alpha \ln(2e\Delta_i^2 n))}{2\Delta_i} + \frac{c_\alpha}{\Delta_i} \right),$$

where

- $\Delta_i = \mu^* - \mu_i$ is the regret associated with choosing arm i over the optimal arm, representing the difference between the maximum expected reward across all arms (μ^*) and the expected reward of arm i (μ_i).
- μ_i is the expected reward of arm i.
- μ^* is the maximum expected reward across all arms.
- α is a parameter that adjusts the level of exploration in the UCB2 policy, influencing the balance between exploration and exploitation.
- n is the total number of plays.
- c_α is a constant dependent on α, further affecting the exploration-exploitation balance.

Nevertheless, the introduction of unobserved confounders into the equation complicates the process of minimising regret. As discussed by Bareinboim et al. [131], traditional bandit algorithms do not fully suffice in scenarios involving unobserved confounders. By reinterpreting the multi-armed bandit problem through causal inference, we can utilise both observational and experimental data to optimise rewards, despite the presence of such confounders.

We proceed to discuss an illustrative example adapted from [131], highlighting the complex interplay between causal structure, policy optimisation, and actor intent. The concept of *intent* is defined as follows.

Definition 4.6.1 (Intent [138]). *In a Structural Causal Model (SCM) M, let Π represent the set of all decision-related variables. For any decision-related variable $\Pi_i \in \Pi$, and at any time t, the intended decision $I_{\Pi_i,t}$, which is a function of both observable and unobservable factors, is defined as follows:*

- $I_{\Pi_i,t}$*: The intended decision for variable* Π_i *at time* t*.*
- $i_{\Pi_i,t}$*: The observed decision made by the actor for variable* Π_i *at time* t*, considering the configuration of unobserved confounders.*
- $U_t = u_t$*: The configuration of unobserved confounders at time* t*.*
- $pa(\Pi_i)$*: The set of parent variables that directly influence* Π_i *in the SCM.*
- f_{Π_i}*: The function that maps the current state of* $pa(\Pi_i)$ *and the unobserved confounders* $u_{\Pi_i,t}$ *to the intended decision* $I_{\Pi_i,t}$*.*

Therefore, the intent for Π_i at time t can be formally expressed as:

$$I_{\Pi_i,t} = f_{\Pi_i}(pa(\Pi_i)_t, u_{\Pi_i,t})$$

Consider the healthcare scenario, a domain riddled with intertwined causal relationships. Here, a treatment decision for a patient isn't just an isolated event; it's a culmination of the patient's medical history, their current health status, the medical professional's expertise, and even external variables like the availability of certain medicines or treatments. Further imagine a scenario where a patient has recurrent migraines. A doctor prescribes a particular medication based on its proven effectiveness. However, the patient doesn't find relief. The intent behind the doctor's decision was to alleviate the patient's pain based on historical data. If we merely analyse this situation using traditional RL models, the doctor's decision might appear sub-optimal given the outcome.

What if the patient had informed the doctor about a previously tried medication that didn't work? What if the doctor knew about the patient's aversion to certain medications due to past side effects? Such nuances reveal the intent behind the doctor's decision. It's not just about treating the migraine; it's about ensuring the patient's overall well-being, considering their past experiences and preferences. Recognising this intent can lead to better treatment decisions in the future, understanding the patient's unique needs, and even identifying biases in decision-making processes.

Given this motivation for why intent could be important, we now consider a numerical example.

Example 10 (A Casino Exploitation Scenario)**.** *Envision a scenario where a casino's slot machines are equipped with technology to identify if a gambler is under the influence of alcohol. These machines have the capability to lure inebriated players through the use of flashing lights, exploiting their diminished capacity to perceive manipulations in payout rates. Nonetheless, the casino employs tactics that circumvent standard testing procedures, creating the illusion of compliance with the legal mandate of a* 30% *minimum payout rate.*

In causal inference, our approach involves segmenting the data according to the specific slot machine that a gambler chooses *to play. This acknowledges the relationship between the player's choice and the payout rates, which is further complicated by the variable of intoxication.*

When we apply this refined concept of choice, our analysis reveals that the real payout percentages hover around 15%*.*

4.6.3 MDPs with Unobserved Confounders

MDPs provide a foundation for RL by simplifying real-world problems into factors of states, actions, and rewards. However, this simplification often overlooks complex interactions, especially with regard to unobserved confounders. To address this, we turn to an augmented model known as Markov Decision Processes with Unobserved Confounders (MDPUC) introduced by Zhang and Bareinboim [139]. Please note that the notation here is consistent with the formulation of a SCM (introduced in Definition 2.7.1) where $\boldsymbol{U}$ are the exogenous variables and $\boldsymbol{V}$ are the endogenous variables. Fur-

(a)	$D = 0$		$D = 1$	
	$B = 0$	$B = 1$	$B = 0$	$B = 1$
$X = M_1$	0.10^*	0.50	0.40	0.20^*
$X = M_2$	0.50	0.10^*	0.20^*	0.40

(b)	$p(y \mid X)$	$p(y \mid do(X))$
$X = M_1$	0.15	0.3
$X = M_2$	0.15	0.3

Figure 4.6: The first table (a) presents a breakdown of the payout probabilities for different slot machines in a casino, categorised by the gambler's sobriety level D, the presence of a flashing light B, and the type of slot machine X. The preferred machines of the gamblers, influenced by their intention, are marked with asterisks. The second table (b) contrasts the observed payout probabilities with those under direct intervention on the choice of machines, highlighting the discrepancy and the failure of simplistic randomisation approaches in uncovering legal non-compliance in the presence of hidden confounders like the flashing light. Adapted from source [131].

ther, F represents the structural equations. We use MDP notation (Definition 3.5.3) for actions, states, and rewards.

Definition 4.6.2 (MDP with Unobserved Confounders (MDPUC))**.** *Defined as a Structural Causal Model (SCM) M, a Markov Decision Process with Unobserved Confounders encompasses an action domain A, state space S, and a binary reward system R:*

1. *The discount factor is denoted by $\gamma \in [0, 1)$.*
2. *At any given time-step t, $U^{(t)} \in U$ represents the unobserved confounder.*
3. *The set of observed variables at time t is $V^{(t)} = X^{(t)} \cup R^{(t)} \cup S^{(t)}$, which includes $A^{(t)} \in A$, $R^{(t)} \in R$, and $S^{(t)} \in S$.*
4. *Structural equations for the observed variables are grouped in $F = \{f_a, f_r, f_s\}$, where the next state and reward are determined by $X^t \leftarrow f_x(s^{(t)}, u^{(t)})$, $R^{(t)} \leftarrow f_r(x^{(t)}, s^{(t)}, u^{(t)})$, and $S^{(t)} \leftarrow f_s(x^{(t-1)}, s^{(t-1)}, u^{(t-1)})$.*
5. *The probability distribution over the unobserved (exogenous) variables U is given by $P(u)$.*

The prevailing discourse in RL literature predominantly revolves around the concept of value functions. In this context, we adapt and elaborate on these concepts specifically for the scenario of MDPUC. This adaptation paves the way for us to seamlessly integrate and leverage the wealth of existing RL research and theories.

Definition 4.6.3 (Value Functions in MDPUC)**.** *For a MDPUC characterised by $M\langle \gamma, U, X, Y, S, F, P(u) \rangle$ and a chosen deterministic policy π, we define the state value function under policy π for a state $s^{(t)}$ as:*

$$V^{\pi}(s^{(t)}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x([t,t+k])=\pi} \mid s^{(t)}\right].$$

In parallel, the function for state-action value is:

$$Q^{\pi}(s^{(t)}, x^{(t)}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x^{(k)}, x([t+1,t+k])=\pi} \mid s^{(t)}, x^{(t)}\right].$$

These value functions in RL are typically interpreted as correlating a given state or state-action pair with the expected cumulative rewards over future time steps. Employing these definitions enables the derivation of the renowned Bellman equation and recursive formulas for both state and state-action value functions, as detailed in [64]. This is particularly pertinent in contexts involving unobserved confounders. The foundational basis for these formulations lies in the counterfactual axioms[1] and the Markov property, as described in theorem (C.3.1). While the proofs for related theorems are detailed in the original work by [139], they are not included here for brevity. The reader is directed to the appendix for a more in-depth discussion of the Markovian properties in MDPUC contexts.

In their research, Zhang and Bareinboim [139] further refine a counterfactual-aware MDP algorithm, showcasing its advantages in scenarios sensitive to intentions. This is expanded by Bareinboim and Pearl [140] to include experimental designs. Forney et al. [109] take a deeper dive, demonstrating that counterfactual reasoning can circumvent the pitfalls of naive randomisation when unobserved confounders are involved. Their approach integrates observational and experimental data for enhanced decision-making, focusing on estimating counterfactual quantities empirically. They investigate how combining data from various sources and under different conditions can augment an RL agent's learning process. A key insight from their work is the separation of *seeing* and *doing* in data analysis. They introduce a heuristic variant of Thompson Sampling [141], empirically shown to surpass previous algorithms. This research also extends the earlier example involving gamblers (Example 10), incorporating the effect of flashing lights on the slot machines, resulting in a matrix of scenarios combining sobriety and machine states.

The discussion begins by recognising that interventional measures, denoted as $\mathbb{E}[Y \mid do(X = x)]$, where X is the variable being intervened upon and Y is the outcome of interest, can be equivalently expressed in counterfactual terms. In this context, $Y_{X=x}$, or more compactly Y_x, represents the counterfactual outcome of Y had X been set to x, even if this may not have actually occurred. This allows us to express the expectation of Y given the intervention $do(X = x)$ as the expectation of the counterfactual outcome Y_x:

$$\mathbb{E}[Y_x] = \mathbb{E}[Y_{X=x}]. \tag{4.1}$$

Utilising the law of total probability, we proceed to derive a practical representation:

$$\mathbb{E}[Y_x] = \mathbb{E}[Y_x \mid x_1]P(x_1) + \cdots + \mathbb{E}[Y_x \mid x_K]P(x_K). \tag{4.2}$$

Here, Y_x is understood as the outcome Y in the counterfactual world where X is set to x, and $P(x_k)$ is the probability of the intervention X being set to a particular value x_k within the set of all possible interventions $\{x_1, \ldots, x_K\}$.

[1] Counterfactual axioms provide a set of principles for causal inference in counterfactual scenarios. They include consistency, which relates observed outcomes to potential outcomes under the actual treatment; independence, which asserts the conditional independence of potential outcomes and treatment assignment given covariates; and SUTVA, which assumes no interference between units and a single version of each treatment.

The terms in this expression are categorised as interventional, observational, or counterfactual, contingent upon whether $x = x'$ or $x \neq x'$, respectively. It is important to acknowledge that counterfactual quantities are typically not directly observable in empirical settings. Nevertheless, the intentions of agents provide valuable insights into the decision-making process, potentially uncovering information about hidden confounders. This aspect was previously discussed in the context of the multi-armed bandit (MAB, see Appendix B.1) scenario. By strategically incorporating randomisation into the agents' intentionality, we can facilitate the computation of counterfactual quantities. This approach enables the enrichment of observational data with insights gleaned from interventional data. To maintain focus and brevity in the main body of this thesis, a detailed exploration of this computational process is presented in the appendix.

4.6.4 Challenges and Solutions in Counterfactual Reasoning within RL

Integrating counterfactual reasoning within RL, particularly in advanced models like MDPUCs, presents several challenges. One of the foremost challenges is the off-policy nature of counterfactual evaluations. This arises when the evaluation policy diverges from the policy currently being followed by the agent.

To address this, one approach is the use of importance sampling. This technique helps to align the distribution of the collected data with that of the new policy, mitigating discrepancies and enabling more accurate counterfactual evaluations [142]. Another method is the twin-network approach, as proposed in [143]. This technique employs two parallel networks: one to predict the expected reward of the action taken and another to estimate the potential reward of the counterfactual action. This dual-network structure allows for a comparative analysis of actual and hypothetical scenarios, enhancing the agent's decision-making process. The advancement of these techniques, especially in counterfactual scenarios, addresses the core of **RQ2**, examining the extent to which causal models can improve the sample efficiency and coordination in decentralised learning tasks.

Counterfactual decision-making stands as a burgeoning area in RL, offering significant advantages in complex environments where traditional models like MDPs may be inadequate. By incorporating augmented models and understanding the role of agent intent, we pave the way for the development of more resilient and adaptable agents. These agents are better equipped to make informed decisions in the presence of hidden variables and biases. Future research should focus on developing reliable methods for counterfactual evaluation, exploring the impact of agent intent in decision-making processes, and investigating the implications of these concepts across diverse applications.

4.7 Generalisability and Robustness

Generalisability and robustness, paramount attributes of human intelligence, enable sophisticated inference and decision-making, even in novel environments. These qualities are essential for developing competent RL agents, regardless of deep learning's role in RL [64].

Definition 4.7.1 (Generalisability (adapted from [48]))**.** *An RL agent generalises if its learnt policy, derived from a source environment, can manage situations or tasks that deviate from prior experiences.*

Generalisability ties closely to the principle of *transfer learning*. In this regard, an agent should utilise prior knowledge and apply it to comparable but distinct domains. Causal models, especially SCMs, streamline knowledge transfer across environments with congruent causal architectures [144]. This is evident in various disciplines, with research delving into the concept of *external validity*, which is fundamentally a form of generalisation. Consider a robot, for instance: if it understands the causal processes involved, training it to pick up boxes would mean it should adeptly handle books. This exploration of transportability and external validity in the context of causal RL directly pertains to **RQ1**. By addressing whether existing RL methods exhibit causal understanding, we delve into their capacity to generalise and apply learnt policies to new, varied environments.

Moreover, a prominent notion in RL bolstered by causal inference is *transportability*. This ability allows for the transfer of causal knowledge across different domains, akin to meta-analysis or externally valid studies. The process of *transportability* becomes vital for agents to streamline knowledge acquisition, discovery, and learning [145]. For clarity, consider the following example.

Example 11 (Transportability Between Los Angeles and New York)**.** *Imagine a team of social scientists who have spent years studying the effect of an educational program, represented by X, on employment outcomes, Y, in Los Angeles. They've noticed that age distribution, Z, acts as a confounding factor, perhaps due to different career opportunities available at various life stages. Given the success and insights from their Los Angeles study, they now wish to predict the same program's impact in New York.*

However, they're aware of demographic and cultural differences between the two cities. Specifically, New York might have a different age distribution due to its distinct job market, cultural appeal, and migration patterns. Therefore, transferring knowledge gained from Los Angeles to New York is not straightforward.

Formally, consider the situation where we aim to utilise the knowledge from the Los Angeles experiments to make analogous predictions in New York. Let's denote the distribution in Los Angeles as $P(y \mid do(x))$. Our objective is to estimate $R = P^*(y \mid do(x))$, which represents the cause/effect relationship under a distinct age distribution in New York.

The mechanism that leads to this age difference across the two populations is termed a difference generating factor *and is graphically represented by* ■, *as depicted in Figure 4.7. This factor arises due to a set of* selection variables*, denoted as S. Consequently, we have a causal link described as $S \rightarrow Z$.*

The relationship between the distributions can be encapsulated using the transport formula*:*

$$\begin{aligned} R &= \sum_{s} P^*(y \mid do(x), z) P^*(z) \\ &= \sum_{s} P(y \mid do(x), z) P^*(z) \end{aligned} \tag{4.3}$$

This elegant formula suggests that we can approximate R, an interventional measure, using a drop in *observational distribution $P^*(z \mid do(x), z)$. Essentially, this serves* *to adjust the weight of observations based on the interventional effect observed in a separate domain.*

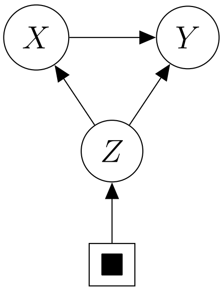


Figure 4.7: A Causal DAG is employed to depict the interconnections among variables X, Y, and Z in the context of social science experiments conducted in Los Angeles. In this graph, the symbol ■ signifies the factor responsible for generating differences, influenced by selection variables S. These variables account for the variations in age distributions observed between the populations of Los Angeles and New York.

4.7.1 Robustness: Standing Steadfast Amidst Uncertainty

Definition 4.7.2 (Robustness)**.** *An RL agent exhibits robustness if it sustains its performance despite uncertainties, model inaccuracies, and possible adversarial disturbances. This trait is also termed local generalisation by, for example, Chollet [48].*

Causal models improve an agent's robustness by enabling it to foresee uncertainties. For example, a causally-informed autonomous car would modify its driving approach during heavy rain, comprehending the causal outcomes of slippery surfaces.

This study posits the following hypothesis, informed by existing literature and logical deduction. This could, for example, suggest a promising direction for future empirical and theoretical research.

Hypothesis 4.7.1. *An RL agent with a grasp on causality is inclined to be both generalisable and robust. This hypothesis draws on the discussion in the preceding text, as well as the associated references, to highlight the potential of causal understanding in enhancing the adaptability and resilience of RL agents across uncertain conditions.*

Additionally, transferring knowledge links well to the big data amalgamation. Merging diverse datasets, gathered under varied conditions without significant bias, is quintessential for reinforcing an agent's capacity to learn in assorted scenarios. This perspective is further considered in Bareinboim and Pearl [140], which scrutinise data fusion through a causal lens, and [146], discussing the extremities of *identifiability* and *randomisation* in discerning cause-effect relationships.

Domain Adaptation in RL: $\mathcal{D}_{\text{training}} \to \mathcal{D}_{\text{testing}}$.

Developing a generalised transport formula is a nuanced task. At its core, the challenge revolves around establishing the completeness of the *do*-calculus: Can the operations of *do*-calculus consistently identify such a transport formula? It's imperative to remember that causal models and their resulting diagrams define relationships within a specific domain.

The challenge of developing a generalised transport formula and the intricacies of domain adaptation in RL resonate with our **RQ2**: Does a causal model improve the sample efficiency and/or coordination of learning agents.

To assist with this challenge, selection diagrams emerged. They visually capture the overlapping causal relationships and the distinct factors that differentiate various causal systems.

Definition 4.7.3 (Selection Diagrams [145]). *Consider a set of two SCMs, denoted as $\langle M, M^* \rangle$, where $\langle , \rangle$ signifies an ordered pair and the first and second elements are SCMs representing domains $\langle \pi, \pi^* \rangle$ respectively, and sharing a common causal diagram G. The ordered pair notation indicates that the models are to be considered together, in the sequence given. These SCMs generate a selection diagram D under the following conditions:*

1. *All edges found in G are replicated in D.*
2. *An additional edge $S_i \rightarrow V_i$ is incorporated into D in instances where disparities are observed between M and M^*, signified by either a variation in functions — $f_i \neq f_i^*$ — or a difference in the probability distributions of the unobserved variables—$P(U_i) \neq P^*(U_i)$.*

These S variables in selection diagrams spotlight the mechanisms where the underlying structural differences manifest between models from disparate domains. Recognising these shared structural elements is foundational to our understanding of knowledge transfer across domains.

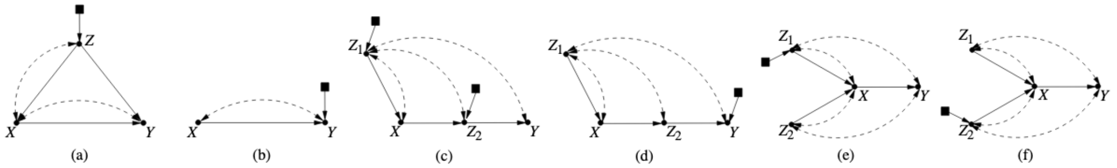


Figure 4.8: Figures (a) to (f) illustrate transportability concepts in causal selection diagrams, emphasising the significance of unobserved confounders. Figure (a) showcases transportability of $R = P^*(y \mid do(x))$ that's efficiently addressed by re-weighting the variable influenced by the difference-generating factor, symbolised by $S \rightarrow Z$. Figure (b) represents a scenario where it's impossible to transfer a causal relationship between domains due to the indeterminacy of the causal effect, even with randomisation on X, because of unidentified confounders (UC). Figures (c) and (d) depict situations necessitating interventional data on Z_1 in π_1 and Z_2 in π_2 for transportability, but not when combined. Meanwhile, figures (e) and (f) illustrate cases where transportability is feasible exclusively within the integrated domain. Adapted from [145].

4.8 Causal Imitation Learning

Bareinboim [107]'s development of the causal RL framework concludes with the exploration of causal imitation learning. This intriguing area dives into imitation learning, where the objective is to glean insights from expert demonstrations. The integration of inverse RL in imitation learning, as investigated by Abbeel and Ng [147], brings us back to the crux of **RQ1**: Do existing RL methods demonstrate a causal understanding? Investigating how causal inference can be embedded within imitation learning frameworks is key to answering this question.

Abbeel and Ng [147] examined how inverse RL (IRL) can be integrated into imitation learning. The essence of IRL lies in deriving a reward function that accentuates trajectories exhibited by experts. This stands in contrast to behaviour cloning, another prevalent imitation learning strategy. Here, an agent attempts to emulate the expert's policy directly. While both approaches have their merits, they both assume that the expert's actions are transparently accessible to the imitator. To address this, the authors present a novel methodology, where a graphical criterion is established to determine when imitation is possible. This determines the feasibility of imitation learning, considering observational data and inherent causal information [148]. In addition, they provide an algorithm to deduce an imitation policy in situations where the set criterion is not met.

Definition 4.8.1 (Partially Observable SCM [148]). *A Partially Observable Structural Causal Model (POSCM) consists of an SCM M, a set of observed endogenous variables $\boldsymbol{O}$, and a set of latent endogenous variables $\boldsymbol{L}$. The union of $\boldsymbol{O}$ and $\boldsymbol{L}$ encompasses all endogenous variables in the model.*

Our focus is on evaluating the effectiveness of a specific intervention, represented by $X \in \boldsymbol{O}$. Considering that the rewards are latent, we aim to identify a policy π that yields an expected reward, $\mathbb{E}[Y \mid do(\pi)]$, exceeding a predetermined threshold τ.

The identifiability of $P(y \mid do(\pi))$ hinges on certain conditions pertaining to exogenous variables, allowing its unique computation from observational data and the POSCM, M. Specifically, the outcomes should be discernible from the observations of expert behaviour in imitation learning scenarios. Nonetheless, if the reward $\boldsymbol{Y}$ is latent, $P(\boldsymbol{y} \mid do(\pi))$ remains unidentifiable without supplemental data to shape an effective imitation policy [148].

When drawing from expert demonstrations, the issue of non-identifiability is diminished. The concept of imitability is thus introduced to describe the replication of a reward distribution $P(\boldsymbol{y})$ based on a pre-established policy within a given policy space and a specific POSCM.

Definition 4.8.2 (Imitation Backdoor [148]). *In a causal system G with a defined policy space Π, a set $\boldsymbol{Z}$ satisfies the imitation backdoor criterion (i-backdoor) relative to $\langle G, \Pi \rangle$ if it is a subset of the parents of Π and ensures that Y is conditionally independent of X after removing X's outgoing edges and considering $\boldsymbol{Z}$.*

The i-backdoor criterion offers perspectives on the practicality of imitating experts in cases where rewards are not directly observable. The subsequent theorem further clarifies this point.

Theorem 4.8.1 (Imitation by Backdoor [148]). *In a causal framework represented by diagram G and policy space Π, the reward distribution $P(y)$ is imitable within $\langle G, \Pi \rangle$ given the existence of a suitable i-backdoor set $\boldsymbol{Z}$. The corresponding policy can be formulated as $\pi(x \mid pa(\Pi)) = P(x \mid \boldsymbol{z})$.*

While the i-backdoor criterion is insightful, it is not the only determinant for successful expert imitation. Further complexities in the relationship between variables and interventions are established through concepts like the imitation surrogate.

Definition 4.8.3 (Imitation Surrogate [149])**.** *For a causal diagram G and a policy space Π, a selected subset of observations $\boldsymbol{O}$, referred to as $\boldsymbol{S}$, constitutes an imitation surrogate (i-surrogate) in the context of $\langle G, \Pi \rangle$ if $(Y \perp\!\!\!\perp \hat{X} \mid \boldsymbol{S})_{G \cup \Pi}$ holds true. Here, $G \cup \Pi$ indicates the graph expanded by directed edges from $Pa(\Pi)$ to X, with $\hat{X}$ representing the newly added parent of X.*

The investigation of i-surrogates and i-backdoors by Zhang et al. provides valuable insights into the prerequisites for policy imitation. Nevertheless, fully capturing all decision-influencing factors in real-world situations remains challenging, making causal imitation learning an ever-evolving field. The pursuit of advancement in this area is a key focus among researchers, given its significant implications for the development of robust and human-aligned artificial intelligence.

The evolution of causal imitation learning and its implications for the development of AI align with the central theme of **RQ2**. It probes into the efficiency and coordination enhancements that causal models can bring to RL agents, especially in complex imitation scenarios.

4.9 Causal Structure Learning and RL

RL empowers agents to devise strategies through interaction with their environments. A fundamental challenge in RL is discerning real cause-and-effect relationships from those confounded by external factors. Incorrectly identified causal relationships can lead to ineffective strategies, particularly in novel situations [150]. This section considers the intricacies of learning causal structures within RL.

Deep RL has demonstrated efficacy in complex environments. However, its “black box” nature often obscures the learning mechanisms, presenting challenges in causal understanding [151]. Integrating causal knowledge with RL promises to yield agents capable of making decisions that are robust against confounding factors, yet this integration is computationally demanding, as detailed in Section 2.11.

The integration of causal knowledge into RL, especially in deciphering cause-and-effect relationships, directly addresses **RQ1**. It questions the current capabilities of RL methods in terms of their causal understanding and the effectiveness of their decision-making processes.

The task of understanding causal relationships has been extensively discussed in both ML and social sciences [152, 153]. Significant progress has been made with algorithms like IC, PC, and GES, which facilitate causal structure identification from observed data, albeit with assumptions such as causal sufficiency. A notable advancement is the algorithm by Kocaoglu et al. [149], which delineates causal graphs and identifies hidden variables with minimal interventions.

RL’s potential to expedite the discovery of causal relationships is an area ripe for exploration [154]. RL agents could discern true causal connections from mere correlations, addressing challenges inherent in traditional causal discovery methods. This synergy between RL and causal inference could lead to a more profound understanding of complex systems.

Kocaoglu et al. [149] enhanced causal discovery by devising an algorithm that identifies any causal graph and determines latent variables with a limited number of interventions. This algorithm operates in three phases: identifying the transitive closure of the observable graph, reducing it to highlight key causal edges, and using conditional independence tests to reveal latent variables. This methodology represents a significant step in causal structure learning, combining computational efficiency with probabilistic rigor.

The potential role of RL in facilitating causal discovery, as discussed, relates to **RQ2**. It highlights how causal models, augmented with RL techniques, could improve the sample efficiency and coordination of learning agents in discovering complex causal structures.

Given the detailed nature of this algorithm and its specific stages, a comprehensive exposition is provided in the appendix for interested readers. This relocation allows the main text to maintain focus on the broader implications of integrating causality with RL, aligning with the thesis's primary argument.

4.10 From Single-Agent to Multi-Agent Realities

Transitioning from the exploration of causal structure learning in RL, we now shift our focus to a broader related domain: multi-agent systems (MAS). This shift is crucial, as the real-world applicability of RL often transcends the confines of single-agent scenarios, venturing into the complex dynamics of multi-agent interactions. The principles of causality and learning, discussed in the context of single-agent models, face new challenges in the context of MAS. This shift to a more complex, multi-agent framework underlines the core of our **RQ1**: Do existing RL methods exhibit causal understanding? In terms of MAS, the need for robust causal comprehension becomes even more critical to navigate the intricate dynamics of agent interactions.

In the traditional RL framework, the focus is typically on a solitary agent learning to optimise actions to maximise cumulative rewards over time. However, this single-agent model simplifies the complexities inherent in many real-world situations, where multiple agents interact simultaneously. In reality, the world functions as a multi-agent system.

Humans, for instance, do not exist in isolation. They continuously interact with other humans, animals, and systems, all of which can be viewed as agents. This multi-agent perspective extends beyond cases with distinct individual entities. Consider, for example, a multi-cellular organism: while perceived as a single entity, it comprises numerous individual cells. These cells, or sub-agents, pursue their specific objectives (such as reproduction or energy acquisition) while collectively contributing to the organism's overall survival and propagation. This cooperation among sub-agents mirrors the dynamics within MAS.

When applying these concepts to RL, mastering an environment means understanding and navigating through these multi-agent systems. The presence of multiple agents, each with its own goals and learning mechanisms, introduces complexities absent in single-agent scenarios. This chapter will discuss these complexities, particularly focusing on integrating causality in MAS, drawing on Pearl's graphical formulation.

Example 12 (Robotic Soccer)**.** *Consider a robotic soccer game where multiple robots interact on the field, each performing distinct actions like kicking, passing, or blocking. In such a setting, the joint state-action space expands exponentially. Employing causal discovery methods, as discussed earlier, could aid in discerning the causal relationships between actions, helping each robot attribute outcomes to specific actions amidst the chaos of other agents. For instance, as one robot improves its passing skill, another robot's interception strategy might need updating. Through causal models, robots can track environmental changes and adjust their strategies in response to their peers' evolving tactics.*

Moreover, a robot must make strategic decisions like whether to pass the ball or attempt a goal. By applying causal reasoning, informed by Pearl's work on SCMs, an agent can contemplate counterfactual scenarios: "If I pass the ball, what are the chances of teammate X scoring a goal?" This level of sophisticated behaviour likely requires understanding and predicting others' actions based on their mental states.

Imagine now, this robotic soccer team is introduced to a new field or faces unfamiliar opponents. Identifying which aspects of their policy depended on the specificities of the previous environment becomes critical. We posit that agentic causal understanding would allow agents to distinguish transferable knowledge from context-specific strategies, adapting their actions to better suit the new environment.

In light of this backdrop, a pertinent question arises: How can we effectively transition towards a causal multi-agent formulation? The subsequent sections aim to address this, exploring the methodologies and implications of integrating causal reasoning within multi-agent environments.

4.10.1 Toward a Causal Formulation of MARL

In the previous chapter, we explored how SCMs apply to single-agent causal RL. Moving to the multi-agent setting introduces new complexities. Here, each agent operates under its own policy, creating a non-stationary environment with dynamics affected by various agents' strategies. This necessitates a sophisticated approach, potentially incorporating counterfactual reasoning to analyse causal interactions.

In environments populated by multiple agents, each with its own policy and objectives, non-stationarity is common. The dynamics of the environment are affected not only by individual agents' actions and environmental randomness but also by the changing strategies of other agents. For example, consider two agents in a shared grid-world. If Agent A notices a change, it's important to determine whether this is due to Agent B's actions or an external factor. Counterfactual reasoning helps here: it allows agents to consider scenarios like, 'If Agent B had not taken action X, would this change still have occurred?'

This approach is especially useful in environments where agents have limited information. With only a partial view of the world, agents need to distinguish changes caused by hidden environmental factors from those due to other agents' strategies. This relates to the 'theory of mind,' where understanding and predicting others' intentions and beliefs is key to forming effective strategies. Counterfactual reasoning enables agents to model their peers' behaviour and intentions, helping them anticipate actions and adapt their strategies.

Therefore, we suggest that a *Multi-Agent Causal Model (MACM)* could be as impactful in MARL as causal perspectives have been in single-agent RL. MACM provides a framework for agents to understand not only their own actions and effects but also the complex interactions between multiple agents. We define this concept in Definition 4.10.2. Our investigation into this area, though ambitious, was more focused. The methods and results of this investigation are detailed in Section 5.1.

4.10.2 Bayesian Perspectives and Factored Models

The shift from causal RL to multi-agent systems brings us to the Bayesian approach. Historically, Bayesian methods have been key in developing our understanding of decision-making under uncertainty. These methods have laid the groundwork for the causal theories that are now central to RL.

Bayesian methods are particularly useful in multi-agent settings where agents often have limited information. They provide a way to think about how agents interact, forming a basis for later developing causal models. In environments where agents do not fully know the strategies or intentions of others, Bayesian models help in managing this uncertainty [155]. The complexity of multi-agent decision-making has led to the use of more manageable models like factored Dec-POMDPs [156]. These models simplify the problem by dividing the state and action spaces into smaller parts, making calculations more efficient.

In this context, Bayesian games have become a useful framework. They model situations where each agent makes decisions based on their beliefs about other agents and their observed actions. This concept of making strategic decisions with incomplete information is a precursor to the more complex causal models in multi-agent systems.

Definition 4.10.1 (Bayesian Game)**.** *A Bayesian game is described as a tuple $\langle N, T, A, u, \pi \rangle$, where:*

- *N is the set of players.*
- *T is the set of potential types for each player.*
- *A includes the actions players can take.*
- *u is the utility function, based on an agent's type and actions.*
- *π describes the probability distribution over types.*

Example 13 (The Lemon Car Game)**.** *Imagine a used car market with two types of cars: good cars termed as "cherries" and bad cars termed as "lemons". Sellers are aware of their car's type, but buyers can't distinguish between them by mere inspection.*

Players N**:** *Two players: Seller (S) and Buyer (B).*

Types T**:** *S has types $T_S = \{lemon,\ cherry\}$.*
B has types $T_B = \{believes\_lemon,\ believes\_cherry\}$.

Actions A**:** *S can either sell or not_ sell.*
B can either buy or not_ buy.

Utilities u**:** *S prefers to sell if car is a "lemon" and may not sell a "cherry" if the price is low.*
B will decide based on their belief about the car's type and the offered price.

Beliefs π**:** *S knows the distribution: $\pi_S(lemon) = 0.6$, $\pi_S(cherry) = 0.4$.*
B believes based on past experiences: $\pi_B(believes\_lemon) = 0.7$, $\pi_B(believes\_cherry) = 0.3$.

The game showcases strategic decisions made under conditions of asymmetric information. If buyers predominantly believe cars in the market are lemons, they might offer a price only lemons are worth, leading sellers of cherries to exit the market—a phenomenon known as "adverse selection".

Moving from Bayesian games to factored models in multi-agent systems, like factored MDPs and Dec-POMDPs, marks a shift from basic agent interaction understanding to tackling more complex environments. Bayesian methods have built the foundation, but causal frameworks go deeper into understanding agent interactions. The next section on Multi-Agent Causal Models will show how causal thinking adds depth to our understanding of these interactions in multi-agent environments. The integration of Bayesian approaches with causal models in MAS settings highlights both **RQ1** and **RQ2**. It questions the extent to which RL methods, enriched with causal Bayesian insights, can achieve a deeper understanding (RQ1) and improved coordination and efficiency (RQ2) in multi-agent domains.

Multi-Agent Causal Models (MACM) MACM serves as a robust framework for decentralised systems. It emphasises interactions via shared and private variables, as depicted in Figure 4.9. By focusing on causal relationships rather than explicit learning processes, MACMs hold promise for counterfactual reasoning, intervention analysis, and understanding emergent behaviours in MAS.

Definition 4.10.2 (Multi-Agent Causal Models)**.** *A MACM encompasses n agents, each characterised by a semi-Markovian model $M_i = \langle V_{M_i}, G_{M_i}, P(V_{M_i}), K_{M_i}\rangle$, where $i \in \{1, \ldots, n\}$. In this definition:*

- *V_{M_i} demarcates the model variables for agent i.*
- *G_{M_i} stands for its causal DAG.*
- *$P(V_{M_i})$ specifies the joint probability distribution over the variables.*
- *K_{M_i} highlights variables shared with other agents.*

Example 14 (Pollution Control in Two Cities)**.** *Consider two neighbouring cities, City A and City B, as illustrated in Figure 4.9. Both cities are industrialised, and their factories contribute to the pollution in the atmosphere. City A has the capability to invest in cleaner technology for its factories, while City B can introduce stricter regulations. Both interventions can reduce pollution. The goal for both cities is to ensure a minimal pollution level for the benefit of their citizens.*

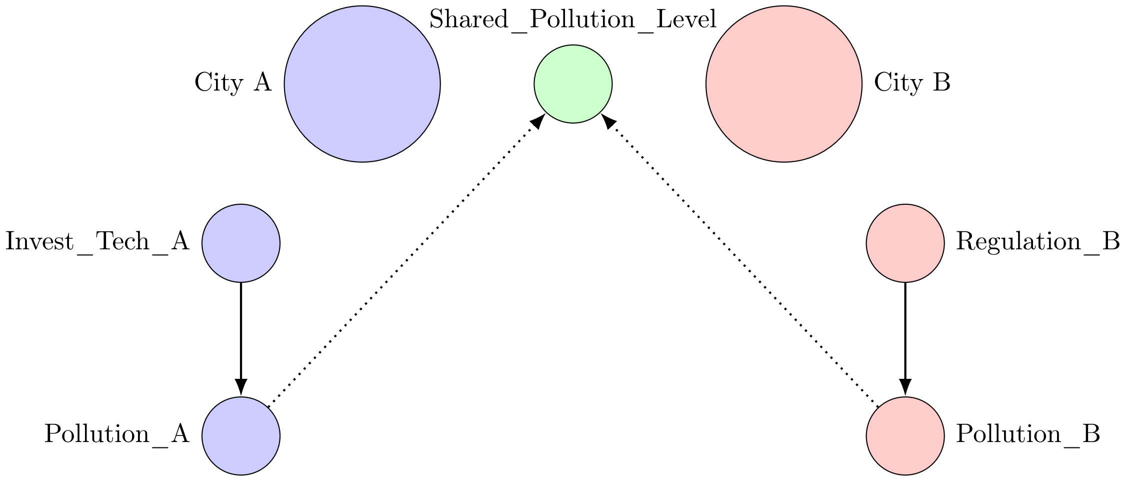


Figure 4.9: A causal DAG with Venn diagram illustrating the relationship between City A and City B in terms of pollution control interventions and their effects. The shared pollution level captures the interdependencies between the two cities.

Agents: *Two agents: City A* (M_1) *and City B* (M_2)*.*

Model Variables: $V_{M_1} = \{Invest\_Tech\_A,\ Pollution\_A\}$.
$V_{M_2} = \{Regulation\_B,\ Pollution\_B\}$.

Causal DAGs: *For City A, if Invest_ Tech_ A increases (representing more investment in clean technology), Pollution_ A decreases.*
For City B, if Regulation_ B is stricter, Pollution_ B decreases.

Joint Probability Distributions: $P(V_{M_1})$ *represents the joint probability distribution of Invest_ Tech_ A and Pollution_ A.*
$P(V_{M_2})$ *depicts the joint probability distribution of Regulation_ B and Pollution_ B.*

Shared Variables: $K_{M_1} = \{Shared\_Pollution\_Level\}$ *- the pollution level affecting City A due to City B's activities. This shared pollution level can be seen at the intersection of the two cities in the Venn diagram.*
$K_{M_2} = \{Shared\_Pollution\_Level\}$ *- the pollution level affecting City B due to City A's activities.*

This MACM example highlights how different interventions in one city can influence pollution levels in the neighbouring city. It highlights the significance of understanding causal relationships in decentralised systems, especially when interventions in one system can have far-reaching impacts on others.

4.10.3 Causality in MAS

Exploring the potential application of causal inference techniques in MAS suggests agents could gain a deeper understanding of the effects of their actions, not only on their immediate environment but also on other agents. While promising, this application of SCMs and Pearl's do-calculus in MAS remains theoretical and speculative. Current research in causal inference provides foundational concepts that could inform this exploration, although its practical realisation in MAS is an emerging area of study [157].

The interdependent nature of MAS, characterised by continuous agent interactions, highlights the value of causal inference. This approach could clarify complex dynamics that are not immediately apparent. For instance, the use of RL in single-agent scenarios has been effective for learning from interventional information [154]. This raises a question related to **RQ1**: can RL's efficiency in single-agent domains extend to uncovering and understanding causal structures in MAS? Such an extension could be groundbreaking, allowing for collaborative learning and adaptation among agents.

Envision a scenario where agents in a MARL setup collaboratively learn the causal dynamics of their environment. They would adapt their models based on both personal experiences and observations of other agents' actions, similar to jointly solving a puzzle. This collaborative process represents an innovative direction in MAS research, blending MARL with causal discovery. While algorithms for identifying MACMs exist [98], their integration with MARL is still in its infancy and presents an exciting area for future research.

Integrating Bayesian methods and factored models with MARL and causal inference could offer more immediate benefits. Bayesian methods, known for their robust probabilistic reasoning, can help manage uncertainty in the causal discovery process. Factored models simplify the learning process by breaking down complex environments into more manageable components. Combining these approaches with MARL and causal inference suggests a novel pathway for understanding MAS, albeit one that requires further exploration and empirical validation. This novel pathway for understanding MAS, while still requiring exploration, directly relates to **RQ2**: *How can a causal model improve the sample efficiency and/or coordination of learning agents in multi-agent environments?*

4.10.4 The Landscape of Multi-Agent Systems and MARL

The study of MAS encompasses a wide range of research areas that have evolved over decades. In this thesis, as we consider MARL, it is crucial to position our discussion within this broader MAS context to appreciate its full scope and identify areas where MARL intersects with other MAS themes.

Themes in MAS such as communication protocols, agent architectures, and coordination mechanisms have significant implications for MARL. Table 4.1 summarises these intersections, highlighting how established MAS concepts can inform and enhance MARL strategies. For example, leveraging communication protocols in MARL could improve agent coordination and learning efficiency. Similarly, understanding distributed problem-solving and constraint reasoning in traditional MAS settings could provide insights into optimising MARL approaches.

The interplay between traditional MAS research and MARL offers fertile ground for future exploration. MARL, with its focus on algorithmic learning and optimisation, gains context and depth when considered within the broader framework of MAS. This convergence of MAS and MARL themes, in relation to **RQ1** and **RQ2**, showcases the potential of causal understanding and its role in enhancing coordination and efficiency within MAS.

Table 4.1: Interplay Between MAS Themes and MARL

MAS Theme	Relation to MARL
Communication Protocols	Agent-to-agent communication modelled as an auxiliary task, enhancing learning and coordination.
Agent Architectures	Leveraging architectures like BDI to guide the exploration-exploitation trade-off in MARL.
Negotiations and Auctions	MARL benefiting from mechanisms by learning optimal bidding or negotiation strategies.
Coordination & Cooperation	MARL providing a framework where agents learn to cooperate or compete in dynamic environments.
Distributed Problem Solving	Agents in MARL learning complementary strategies to solve sub-components of larger challenges.
Constraint Reasoning	MARL employed to dynamically learn constraints, adjusting agent behaviours accordingly.
Swarm Intelligence	Simple rules for emergent behaviours - global objectives are met via local interactions.
Robustness & Fault Tolerance	Agents in MARL trained to adjust strategies dynamically in the face of agent failures.
Environment Modelling	Agents in MARL learn and adapt to environmental nuances, capturing agent-environment dynamics.

4.11 Conclusion

This comprehensive literature outlined in this chapter has provided insights into the current state of research at the intersection of causality and RL. This review has contributed to identifying key challenges and emerging solutions within the field. The findings emphasise the role of causal methodologies in advancing RL, highlighting their potential to address issues such as generalisation, sample efficiency, and multi-agent coordination.

Key takeaways from the review include:

- **Graphical Representations and Counterfactual Reasoning**: These have been identified as essential tools for integrating causality into RL. Their application could lead to a better understanding and enhanced performance in complex environments.
- **Research Direction and Clarity**: The review revealed some ambiguity in the research community's objectives regarding the role of causality in RL. There is a need for clearer goals: whether the aim is to use causality to improve RL or to use RL as a means to deepen causal understanding.
- **Future Research Potential**: The merging of causality with RL, particularly in developing causal model-based methods, offers significant opportunities. Fo-

cusing on the causal aspects of RL methods could advance the development of sophisticated agents and improve learning efficiency.

- **Untapped Avenues**: Certain ideas from the early stages of this research remain underexplored. While not the primary focus of this study, these topics have shaped our overall research approach and offer potential for future exploration.

These findings underscore the importance of causal methodologies in addressing two of three research questions — specifically **RQ1** and **RQ2**. The findings highlight how causal thinking can enhance RL, particularly in addressing challenges like generalisation, sample efficiency, and coordination in multi-agent settings.

The next chapter will investigate causal multi-agent RL. We will apply the knowledge gained from the literature review to examine focused studies in this area. Additionally, we introduce and investigate potential biases that arise in cases where learnt causal models are applied in practice, particularly concerning fairness in ML applications. This discussion is important as it emphasises the need to balance advanced technical approaches with ethical considerations.

Chapter 5

Methodology and Results

This chapter presents the research methodology and integrates key results where relevant. It focuses on an extensive literature review and theoretical development, presenting results alongside the methodology for coherence.

Section 5.2 outlines the methodology and outcomes of a specific investigation. While not all avenues are detailed, two areas are emphasised: (1) the literature review and theoretical investigations, and (2) the analysis of induced disparities in subgroups during causal discovery. The latter, empirical and practical, addresses **RQ3** - *Can applying a learned causal model in causal inference lead to disparate impacts on sensitive subgroups?*. Methodology and results are co-presented under structured subheadings.

Following the literature review in Chapter 4, this chapter focuses on two main studies: (1) enhancing Multi-Agent Reinforcement Learning with causal inference, and (2) using causality to address disparate impacts.

The first study explores combining causal inference with causal RL in Multi-Agent RL, examining potential enhancements. The second investigates causality in reducing disparate impacts, especially in algorithmic decision-making where causal models are learned. These areas were chosen for their relevance and potential impact on the thesis's overarching theme. They build upon foundational concepts from the previous chapter, leading to an in-depth exploration of causality and RL.

5.1 Enhancing Multi-Agent RL with Causal Inference

This section considers how causal inference principles can enhance methods and models in Multi-Agent Reinforcement Learning (MARL). The findings of this exploration were published in Grimbly et al. [2].

5.1.1 Methodological Foundations

The core of this research involved extending ideas presented in the extended literature review on causal RL (Chapter 4) to the multi-agent RL setting. This integration introduced theoretical enhancements to the conventional Markov Decision Process (MDP) formulation in several ways:

- **Data Fusion:** Our exploration into data fusion centered on explicating assumptions about the data-generating system and discerning causal relationships between key variables. This approach is based on the premise that understanding these relationships enables the merging of datasets collected under diverse conditions and policies. This concept aligns with the research by Marcellesi [158] and Bareinboim and Pearl [159], who have demonstrated methods and theoretical frameworks for effective data fusion in causal inference.

- **Off-Policy and Offline Learning:** The integration of causal inference tools into RL has been particularly effective in off-policy and offline learning scenarios. These tools provide mechanisms for bias correction when learning from preexisting datasets, thereby enhancing the scalability of RL. This aligns with the findings of Levine et al. [124] and the instrumental variables approach to causality discussed by Becker [160], both of which are important for understanding and addressing biases in RL.

- **Counterfactual Reasoning:** We investigated the application of counterfactual reasoning to improve the data-efficiency of RL algorithms. This involves the interrogation of 'what-if' scenarios, which is crucial in contexts where actions are costly or risky. Our approach is informed by the work of Bareinboim et al. [131] who emphasise the value of counterfactual reasoning in RL and explore this in the context of decision-making under uncertainty.

- **Causal Learning:** A significant part of our methodology involved discussing causal structure discovery from data, particularly methods that test for conditional independence, along with considering factors like time and noise. This aspect of causal learning is supported by the works of Glymour et al. [161], which provide foundational methods in causal discovery, and Löwe et al. [162], who focus on the specifics of discovering causal relationships in time-series data.

This investigation addresses **RQ1**: *How can causal inference principles enhance RL methods and models?* and **RQ2**: *What are the implications of integrating causality into RL?*. Extending these concepts to multi-agent RL, we seek to uncover insights and methodologies to improve RL in dynamic environments.

5.1.2 Application to MARL and Results

Focusing on the application in MARL, a key methodological development was the articulation of a Decentralised partially observable Markov decision process (Dec-POMDP) as a multi-agent causal model (MACM).

Two-Agent Scenario with SCMs: In a two-agent setup under the MACM framework, the trajectories of state, observation, and action were conceptualised as structural assignments in a causal context. This perspective led to the development of two separate SCMs, each representing one of the agents. This approach was useful for clarifying the interactions and commonalities between the agents. A key observation was that both agents interacted with the same environment, sharing a global state and a history of state-action trajectories. While this interpretation closely mirrors the SCM-based understanding of Dec-POMDP, our MACM methodology distinctively handled the shared aspects of agent variables, a feature not typically addressed in standard SCM frameworks. This subtle difference was particularly beneficial in situations where the agents' behaviours diverged from the typical Dec-POMDP model, especially in environments where agents collaboratively strive for optimal results without a common reward structure. In these instances, MACMs served as a 'causal wrapper,' providing a structured way to capture the dynamics of multi-agent interactions more comprehensively.

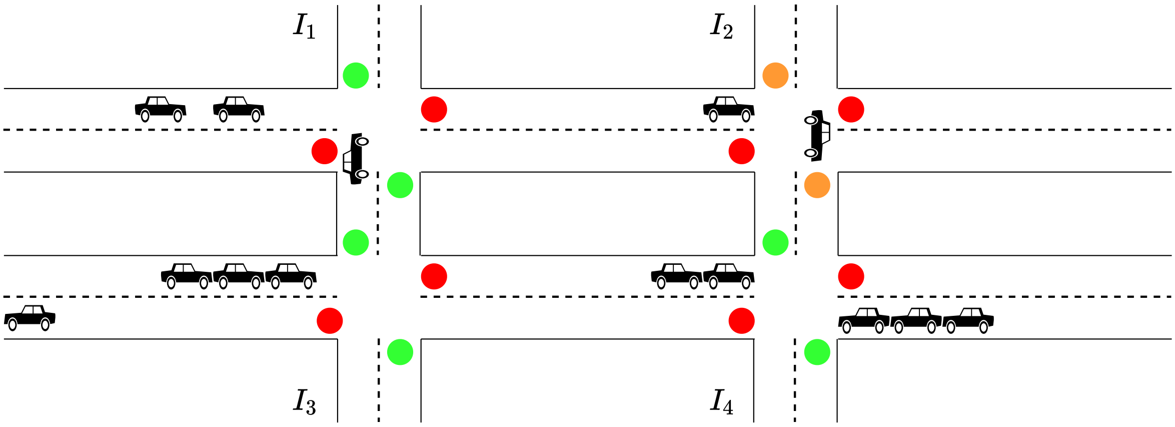


Figure 5.1: Illustration of a network of four intersections, denoted as $\{I_1, \ldots, I_4\}$, conceptualised as individual agents. The primary objective of these agents is to optimise overall traffic flow by managing their respective traffic lights, with their decision-making process based solely on the traffic data from surrounding roads. Optimal viewing in colour.

Expanding to General Cooperative Scenarios: The versatility of MACM was further tested in more complex cooperative scenarios, such as systems of interconnected intersections (Figure 5.1). Here, observations were shared among subsets of agents, demonstrating the flexibility of the MACM approach.

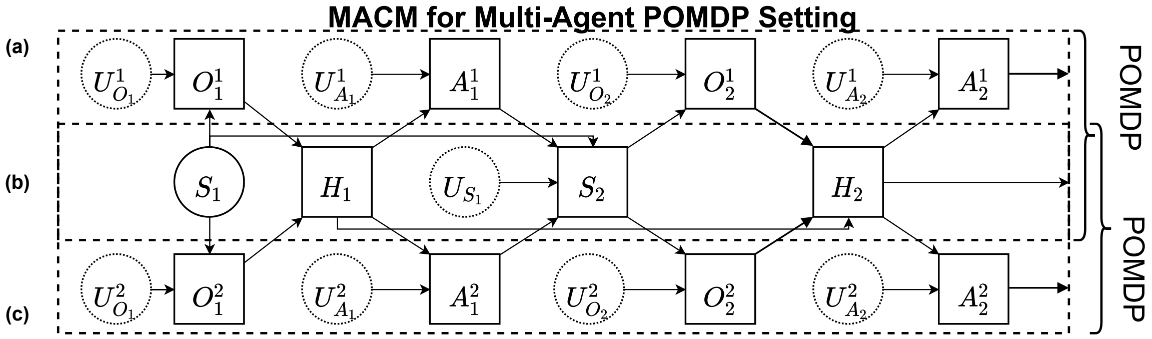


Figure 5.2: This schematic presents a dual-agent system for sequential decision-making, interpreted through a MACM framework. Rectangles in the diagram symbolise endogenous variables linked to their causal assignments, while exogenous variables are shown as circles. Dotted lines indicate the presence of noise variables. In this framework, parts (a) and (c) represent the individual POMDPs of the agents, structured as Bayesian causal networks with explicit inclusion of noise variables. The global state is denoted by S, individual agent observations by O, combined historical data (encompassing past observations, actions, or rewards, subject to the specific scenario) by H, and the agents' actions by A. Agents and noise variables are denoted by superscripts and U, respectively. Segment (b) demonstrates the interconnected areas of the POMDPs. While this diagram is specific to an MACM, the model can also be applied to broader MARL scenarios.

Findings: The methodological innovations in this investigation revealed significant potential for causal inference to augment MARL models and methods. Notably, the 'causal wrapper' provided by MACMs offered a structured way to model agent interactions and shared components in various scenarios. This advancement has opened new avenues for advancing MARL frameworks, particularly in settings where agents collaborate without a shared reward, diverging from traditional Dec-POMDP models.

The development of MACM and its application in various scenarios demonstrate practical advancements in answering RQ1 and RQ2. These findings illustrate the potential of causal inference to enhance MARL, offering structured models that capture the dynamics of agent interactions and shared components, crucial for effective multi-agent coordination and decision-making.

5.1.3 Conclusion

This investigation highlighted the potential impact of causal inference and causal discovery methods in MARL, especially in the context of reformulating Dec-POMDPs as MACMs. By introducing structured and explicit modeling of agent and environment interactions, MACMs offer a potential for significant advancements in the field of MARL.

5.2 Causality for Addressing Disparate Impacts

In an era increasingly dominated by data-driven decision-making, the ethical ramifications of ML models, especially regarding their impact on marginalised communities, have become a focal point of scholarly attention. Aligning with the methodology and results theme of this chapter, this section considers the application of causal inference methodologies in ML to address fairness and bias issues. As discussed in Section 2, causal methods are instrumental in identifying and mitigating biases in ML applications [163–166]. Leveraging pre-existing domain knowledge is fundamental to these methods, providing the necessary groundwork for developing causal models that help rectify biases in data-driven systems.

When domain knowledge is sparse, causal discovery techniques become essential. These methods are designed to extract causal structures directly from data, supplementing limited domain expertise [167–169]. This approach informed our investigation into **RQ3** - *Can learning a causal model and applying it for applied causal inference lead to disparate impacts on sensitive subgroups?*

Our methodology, detailed in this section, aims to be comprehensive and self-contained, reflecting a distinct sub-investigation within the broader scope of this MSc thesis. It encompasses both an exploration of the problem and an evaluation of model fairness, employing quantitative and qualitative methods to assess potential biases in fields such as education and criminal justice. The significance of this evaluation is amplified by the extensive discussion surrounding model fairness and bias in existing literature [170, 171]. Subsequent sections will expound on the specific methods and analytical techniques utilised in our research, illustrating their role in mitigating disparate impacts and fostering equitable outcomes in algorithmic decision-making.

In addressing these biases, a critical aspect is the consideration of **sensitive variables**, such as race, gender, and socioeconomic status. These variables often represent the characteristics of marginalised groups most affected by algorithmic decisions, but one can define a variable as sensitive in an abstract setting. In this sense, a sensitive variable is a variable in the problem setting that has different properties or must be treated differently. Due to lack of specificity here, practitioners would defer to domain experts. In our methodological approach, identifying and appropriately handling these sensitive variables is fundamental to ensuring fairness in causal models.

5.2.1 Problem Setting and Approach

This subsection presents the methodological framework adopted to address **RQ3** - *Can learning a causal model and applying it for applied causal inference lead to disparate impacts on sensitive subgroups?* (see Section 1.1). The methodology is grounded in the ethical imperative to understand and mitigate potential biases against sensitive groups within ML models.

The approach was developed during an internship focusing on bias and fairness in ML, under the supervision of Prof. Ferdinando Fioretto at Syracuse University. Central to our methodological framework was the use of the *Causal Discovery Toolbox* in R [172] and specific Python packages [173–175]. These tools were selected for their robustness in modeling complex causal relationships and their suitability for exploring fairness-related issues in ML.

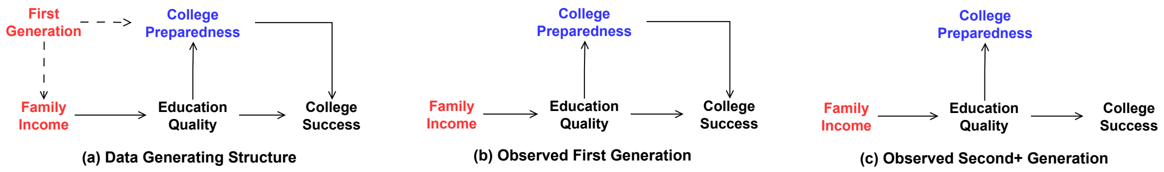
In our methodological design, we prioritised graphical causal models to represent the interactions between variables related to sensitivity and fairness. The choice of these models was informed by their ability to represent intricate dependencies and their applicability to real-world scenarios. This choice was also driven by the need to illustrate how conventional causal models might overlook or misrepresent the dynamics affecting marginalised groups.

A critical aspect of our methodology involved conceptualising and operationalising variables within these models. For instance, we introduced a latent sensitive variable s as an exogenous factor in our DAG, denoted as $s \in \boldsymbol{U}$. This decision was based on our objective to challenge and investigate the *no confounding* assumption prevalent in many structure-learning algorithms (as elaborated in Section 2.11.3). Similarly, we included an observed sensitive variable m as an endogenous factor, represented by $m \in \boldsymbol{V}$, to further probe the complexities of sensitivity within causal models.

Example 15 (Why is this important?)**.** *To illustrate the application of our methodology, consider the example of universities formulating policies to address disparities in college success.*

Imagine that universities are considering implementing new policies that account for disparities in college success by adjusting for family income, similar to an affirmative action policy. By implementing a policy on a causal graph learnt from the entire observational dataset, an unfair policy would be justified.

The unfairness of such a policy is highlighted by the disparate graphs learnt between classes. In Figure 5.3(b) we show an example of a possible learnt graph where all students are first generation. Here the First Generation (as in 5.3(a)) variable is missing because this variable is not recorded in the observational dataset. Note the edge identified for College Preparedness → College Success. This contrasts with Figure 5.3 (c) where the edge is not identified. This occurs because College Preparedness is not (or weakly) important for legacy students in this example.



(a) Data Generating Structure

(b) Observed First Generation

(c) Observed Second+ Generation

Figure 5.3: Figure (a) depicts the foundational data structure for college success, highlighting key variables such as *education quality* and *college preparedness.* This model serves as a baseline for understanding the causal dynamics in educational success and forms the basis for our comparative analysis of bias in different student groups. In this example, education quality and college preparedness are important predictors of college success, with education quality also being a causal predictor of college preparedness. An important predictor of college preparedness is whether or not the student is first generation - the first in their family to attend college. Further, higher education quality results in higher college preparedness. The first generation factor also has a causal influence on family income, as college educated parents are more likely to earn higher salaries.

5.2.2 Fairness Metric

A component of our methodology involves the conceptualisation and application of a fairness metric. This metric is crucial for evaluating the outcomes of our causal models, especially in the context of disparate impacts on sensitive subgroups.

The fairness metric in this study is formulated based on the Bayesian Information Criterion (BIC). The BIC is a standard scoring mechanism in graph learning algorithms such as GES, providing a quantitative measure of the model's *goodness of fit* while penalising model complexity. Mathematically, BIC is expressed as:

$$\mathrm{BIC} = k \ln(n) - 2\ln(\hat{L}) \tag{5.1}$$

where k represents the number of parameters in the model, n is the sample size, and $\hat{L}$ is the maximised value of the likelihood function of the model.

Building upon this, the fairness metric in our study is designed as a bound on the difference between the BIC of subgroup-specific causal graphs and the BIC of causal graphs derived from the entire dataset:

$$\text{Fairness Metric} = |\mathrm{BIC}_{\text{subgroup}} - \mathrm{BIC}_{\text{total}}| \tag{5.2}$$

Here, $\mathrm{BIC}_{\text{subgroup}}$ is the BIC value for the causal graph of a specific subgroup, and $\mathrm{BIC}_{\text{total}}$ is the BIC value for the causal graph based on the entire dataset.

While the BIC-based fairness metric offers a starting point for evaluating fairness in causal models, it is important to acknowledge its limitations. Notably, the BIC fairness metric does not inherently encapsulate causal relationships, making it less ideal for studies prioritising causal interpretations of fairness. However, its use in this research serves as an initial step in understanding how fairness metrics can be integrated into broader fairness assessment and mitigation pipelines.

A more causally-oriented metric, such as the Structural Intervention Distance (SID), could potentially offer a more nuanced understanding of fairness in causal models.

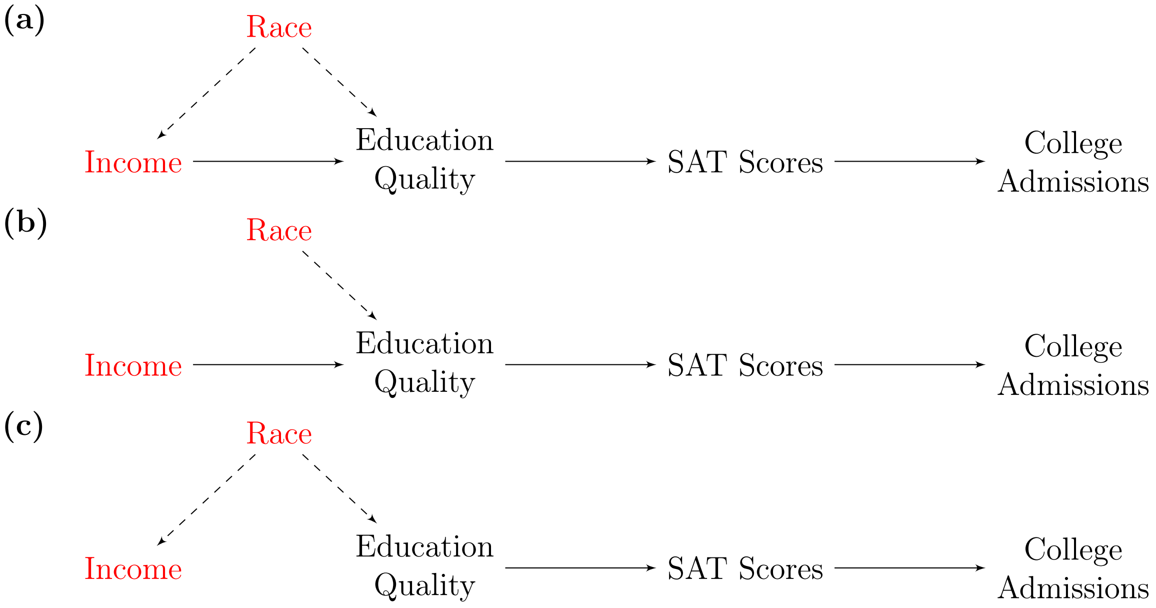


Figure 5.4: Data generating DAG structure for 5 variable model. Here the outcome variable of interest is *College Admission*, which depends on several other variables. Of particular interest are the sensitive variables, *income* and *race* (shown in red). These are sensitive as we do not want the outcome to depend on them. We generate data according to three (seemingly) possible and realistic structures. We show an example of (a) an assumed data structure, (b) a possible different domain where *income* is not *race* dependent, and (c) the structure in (a) after some policy intervention which ensures *education quality* is not directly dependent on income.

SID, which measures an interventional distance between causal structures in terms of the distributions they induce, could provide a more direct assessment of the fairness implications of causal relationships. However, the exploration of SID as a fairness metric represents future work beyond the scope of this study.

In summary, the BIC-based fairness metric used here acts as an initial foray into the quantification of fairness in causal models. It lays the groundwork for future investigations that could incorporate more causally-aligned metrics, enhancing our ability to detect and mitigate biases in ML algorithms.

5.2.3 Data Generation

Inspired by the LUCAS dataset [176], which offers plausible yet artificial relationships between variables, our data generation process aimed to create datasets reflecting defined conditional dependencies. The LUCAS dataset's structure provided a template for simulating realistic, yet controlled, variable interactions, critical for exploring the impact of biases in causal discovery.

5.2.4 Class and Magnitude Imbalancing

Gender, a frequently examined variable in fairness research, was selected for its binary nature and prevalence as a sensitive variable. Our focus was to understand how class imbalancing, especially in binary-sensitive variables like gender, affects the discovery of causal structures. The experimental procedure involved generating and manipulating datasets to reflect imbalances in gender representation, starting with a base size of

5000 samples. The steps included:

1. Constructing datasets with balanced gender representation, ensuring 50% representation for both male and female categories.
2. After incorporating gender's influence on other variables, we removed the gender variable to analyse its indirect effects.
3. Partitioning the dataset into equal male and female samples, creating a balanced baseline for comparative purposes.
4. Introducing gender imbalances by adjusting male and female sample sizes and incrementally varying these imbalances.
5. Replicating this process with different random seeds to generate diverse datasets, simulating multiple training scenarios.
6. Comparing the performance across imbalanced classes to the balanced baseline and evaluating the accuracy against the known causal DAG.
7. Assessing various algorithms, such as GES, LiNGAM, and PC, using metrics like AU-PRC, SHD, and SID.

This experimental setup aims to shed light on the influence of class imbalancing on causal discovery, particularly when sensitive variables like gender are considered. The subsequent section presents the results derived from these procedures, focusing on the effects of gender imbalances on causal structure identification.

Results

The results section delineates the outcomes from the experimental procedures described above. We specifically focus on the impact of gender imbalances on causal structure discovery. Figures 5.7 showcase the results of applying the Greedy-Equivalence Search (GES) algorithm on the artificially generated dataset, with varying proportions of male to female imbalances. The graphs learnt under these conditions were compared against the balanced dataset and evaluated using common metrics such as SHD, SID, and AUPRC.

5.2.5 Real-World Data

Recognising the limitations of artificial datasets, our study extends to the evaluation of real-world datasets commonly referenced in fairness and bias literature within ML. This shift towards actual datasets, like the Adult dataset [177], is crucial to validate our findings in more complex and less controlled environments. The unknown ground truth in these datasets necessitated a comparative approach, where graph distance metrics are evaluated against graphs learnt from balanced datasets.

Results

Similar to our controlled experiments (Figure 5.7), analyses on real-world datasets, such as the Adult dataset, are conducted to observe the impact of gender imbalances. Here,

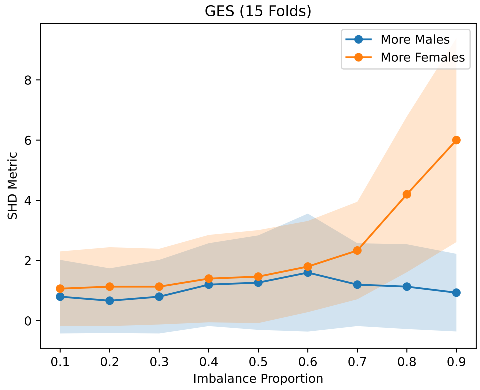


Figure 5.5: LUCAS Dataset - Gender Imbalance SHD Analysis

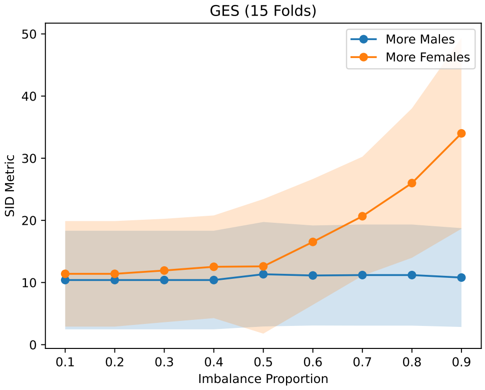
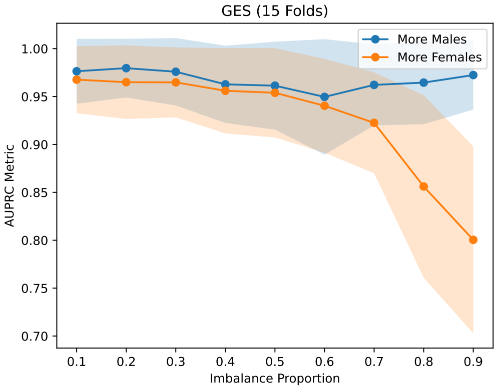


Figure 5.6: LUCAS Dataset - Gender Imbalance SID Analysis



(a) LUCAS Dataset - Gender Imbalance AUPRC Analysis

Figure 5.7: Results for running the GES algorithm on the recreated LUCAS dataset with gender imbalances. Graphs are compared using metrics like SHD, SID, and AUPRC to evaluate the effects of imbalancing male and female data on the accuracy of causal structure discovery.

gender is treated as a latent variable, and findings are compared using common metrics (Figure 5.10). These results underline the potential effects of gender imbalances on causal discovery, reinforcing the need for balanced representation in real-world datasets.

5.2.6 Structural Properties

This section focuses on the effects of graph structural properties on disparate impacts, transcending class and magnitude imbalances. We investigate causal structures with varying configurations of chains, colliders, and forks, generated from distinct ground truth causal DAGs. The positioning of a latent sensitive variable within these structures, particularly in scenarios with unobserved confounding, was a key area of focus.

Many causal discovery algorithms are based on the assumption of no unobserved confounding. Testing these algorithms in environments with hidden confounders, though challenging, mirrors the complexities encountered in real-world data. Our analysis centres on how the scaling behaviour of BIC fairness varies with the shortest directed path between the latent sensitive variable S and the outcome variable Y. The hypothesis

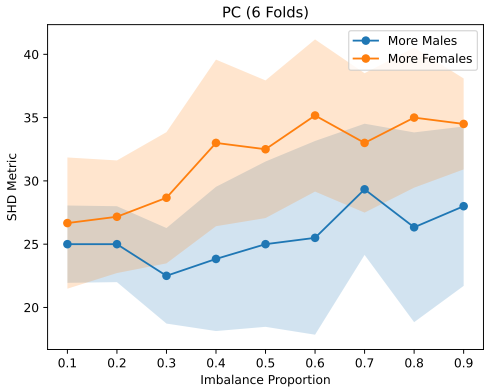


Figure 5.8: Adult Dataset - Gender Imbalance SHD Analysis

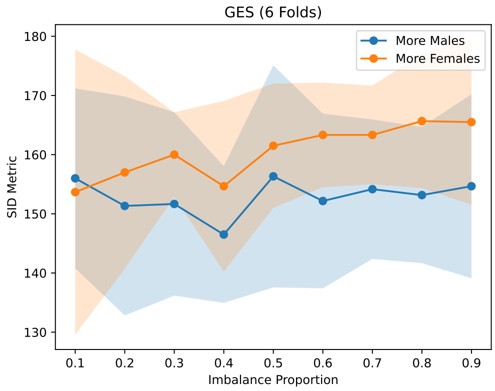
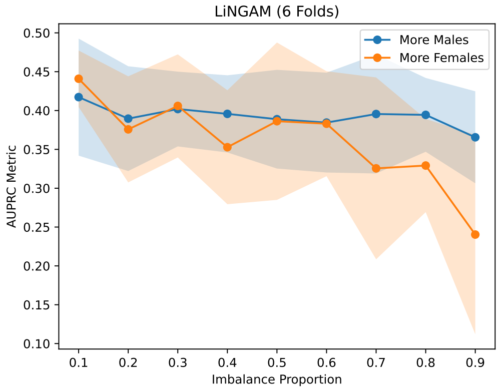


Figure 5.9: Adult Dataset - Gender Imbalance SID Analysis



(a) Adult Dataset - Gender Imbalance AUPRC Analysis

Figure 5.10: Example figure showing sample of results when running common causal discovery algorithms on the Adult dataset, and using common metrics for comparing graphs.

posits that BIC fairness diminishes as S moves closer to Y, due to increased disparity in subgroup and combined dataset outcomes.

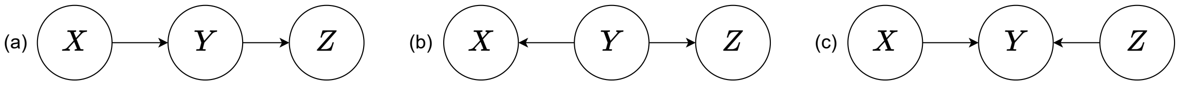


Figure 5.11: Illustrative DAG structures highlighting the importance of structural properties in causal effect identifiability.

Results

Our exploratory analysis reveals how BIC fairness scaling correlates with DAG structures. The results from various scenarios, including simple causal chains and configurations with confounders and forks, provide insights into the nuanced ways these structural elements influence BIC fairness. This exploration is vital for advancing towards algorithmic fairness in complex causal settings, as it unravels the subtle interplay between graph structure and fairness metrics.

Shortest path in a chain. The objective here was to observe how the length of the causal chain impacts BIC fairness when a latent sensitive variable directly influences

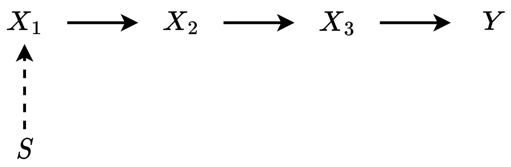


Figure 5.12: Experimental DAG - Simple Chain

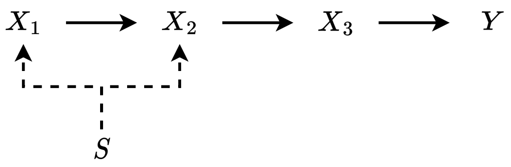
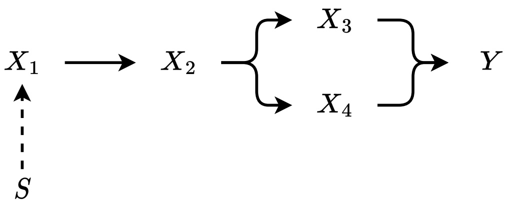


Figure 5.13: Experimental DAG - Chain with Confounder



(a) Experimental DAG - Chain with Fork

Figure 5.14: Experimental structures testing the BIC fairness metric scaling across different DAG configurations, focusing on the path between latent sensitive variable S and outcome variable Y.

one observed variable in a simple directed chain without confounders. Figure 5.17 illustrates the results of this exploration. Figure 5.17(a) depicts the scaling of BIC with chain length, showing how the position of the latent sensitive variable in the causal chain influences the BIC of the learnt model on both the total dataset and on one sensitive subgroup. Figure 5.17(b) highlights the BIC fairness, which is further normalised by the number of samples in the dataset in Figure 5.17(c), displaying how the normalised BIC fairness varies with the chain length.

Shortest path with a confounder. This section examines the impact of introducing a confounder in the causal chain on the scaling behaviour of BIC fairness. We hypothesised that incorporating a confounder would alter the scaling characteristics when compared to a chain without confounding. In this context, the latent sensitive variable influences two observed variables, thereby serving as a confounder. The results of this investigation are depicted in Figure 5.20. Specifically, Figure 5.20(a) demonstrates the BIC scaling with chain length. Notably, Figures 5.20(b) and 5.20(c) reveal the scaling properties for BIC fairness and normalised BIC fairness, illustrating a distinct scaling behaviour due to the presence of the confounder.

Multiple causal paths. This section aimed to understand how multiple causal paths, particularly the inclusion of a fork in the causal chain, influences BIC fairness. In scenarios where the causal structure becomes complex with multiple paths, the latent sensitive variable acting as a confounder results in an increased number of causal paths between variables before and after the fork. Figure 5.23 provides a visual representation of this analysis. Figure 5.23(a) shows the scaling with chain length, while Figure 5.23(b) and Figure 5.23(c) depict the scaling of BIC fairness and normalised BIC fairness respectively. This scenario portrays a different scaling behaviour, shedding light on the interplay between multiple causal pathways and BIC fairness.

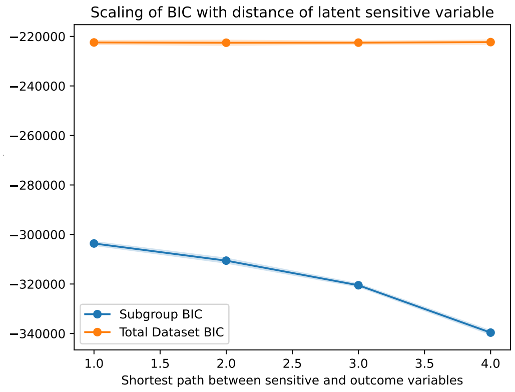


Figure 5.15: Simple Chain - Scaling with Chain Length

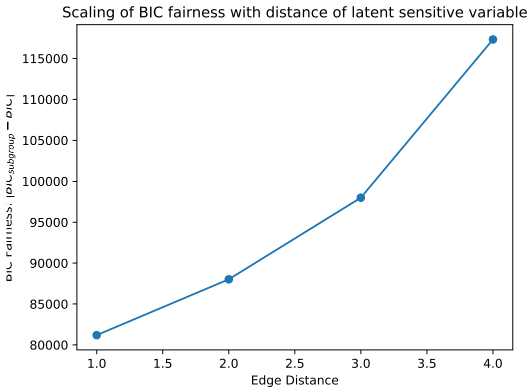
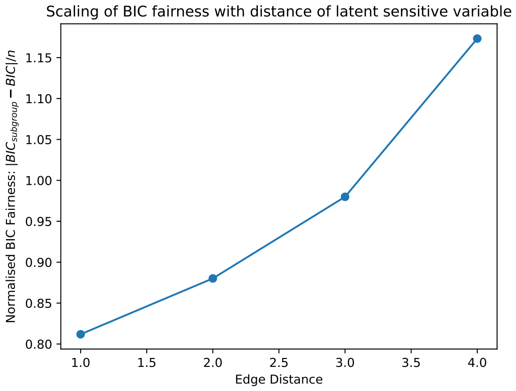


Figure 5.16: Simple Chain - Scaling of BIC Fairness



(a) Simple Chain - Normalised BIC Fairness

Figure 5.17: Results from experiments where a latent sensitive variable directly influences one observed variable in a simple directed chain (no confounders). In (a) we plot the BIC of the learnt model on both the total dataset and on one sensitive subgroup. The absolute difference gives the BIC fairness in (b), which is normalised by the number of samples in the dataset in (c).

5.2.7 Conclusion

This section has examined the use of causal learning methodologies in ML applications, primarily to investigate and address bias and fairness issues. We emphasised the importance of domain knowledge in creating causal models and the relevance of causal discovery techniques where such knowledge is limited. Our focus was on exploring the structural and functional properties of learnt causal models and their implications. By employing graphical causal models, we gained insights into the relationships between variables and how biases might be mitigated in algorithmic systems.

The implementation of a fairness metric based on the BIC was a notable aspect of our study, providing an initial framework for fairness evaluation. However, the metric's limitations suggest the need for further research into more refined metrics for a deeper understanding of fairness.

The experiments with class and magnitude imbalances, particularly using gender as an illustrative variable, highlighted the effects of such imbalances on causal structure discovery, and by extension, real-world impacts such methods could induce. Addition-

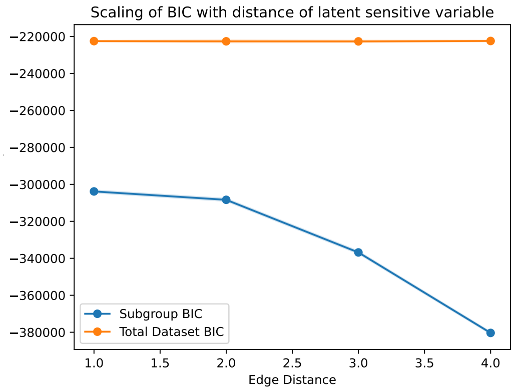


Figure 5.18: Confounder Chain - Scaling with Chain Length

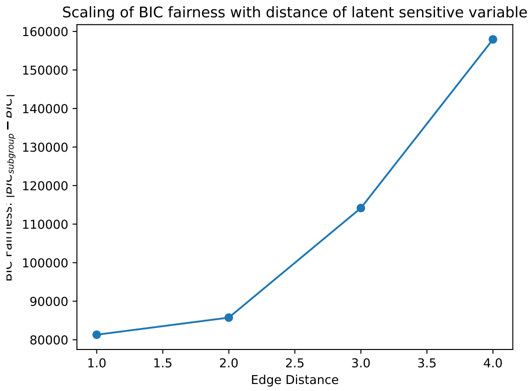
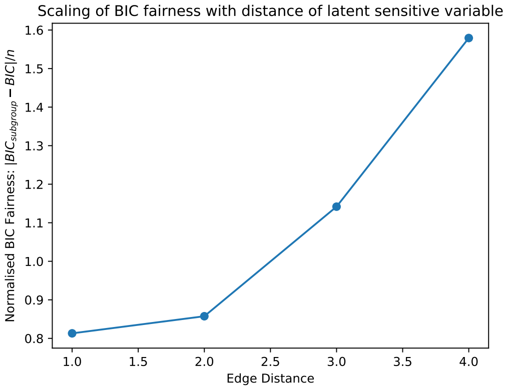


Figure 5.19: Confounder Chain - Scaling of BIC Fairness



(a) Confounder Chain - Normalised BIC Fairness

Figure 5.20: Results from experiments where a latent sensitive variable directly influences two observed variables in a simple chain. In this case, the latent sensitive variable acts as a confounder. In (a) we plot the BIC of the learnt model on both the total dataset and on one sensitive subgroup. The absolute difference gives the BIC fairness in (b), which is normalised by the number of samples in the dataset in (c). We notice a sharper scaling property as opposed to the case where the latent variable is not a confounder.

ally, the exploration of graph structures revealed their influence on fairness outcomes in ML algorithms.

In summary, this investigation highlights the potential of causal inference in ML, but presents evidence for caution in applying causal discovery techniques as a drop-in replacement of prior domain knowledge. This points toward the necessity for ongoing research in this evolving field.

5.3 Conclusion

This chapter has presented a comprehensive overview of the methodologies employed in this research, focusing on the intersection of causal inference and RL, as well as an investigation into fairness, and bias in ML. We began with a detailed exposition of our data generation process, drawing inspiration from the LUCAS dataset, to create realistic yet controlled environments for our experiments. The exploration of class and

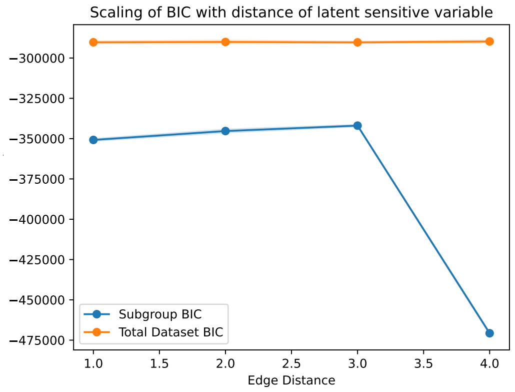


Figure 5.21: Multiple Paths - Scaling with Chain Length

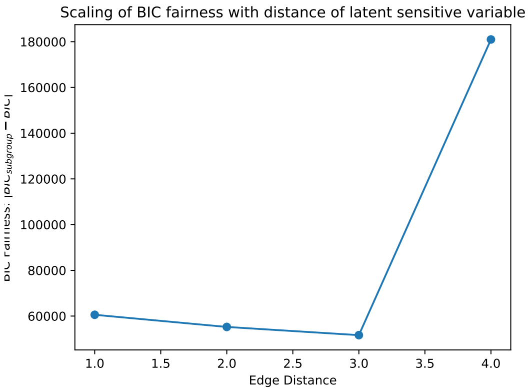
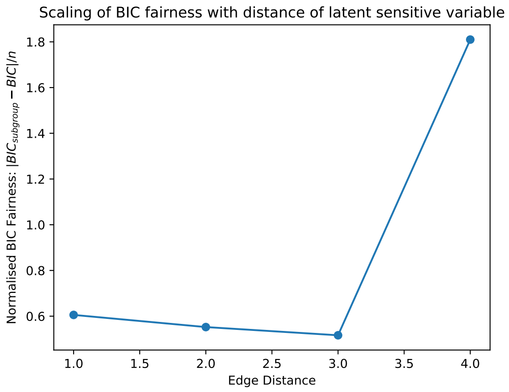


Figure 5.22: Multiple Paths - Scaling of BIC Fairness



(a) Multiple Paths - Normalised BIC Fairness

Figure 5.23: Results from experiments where a latent sensitive variable directly influences multiple observed variables in a complex chain. This configuration explores the effects of multiple paths on BIC fairness metrics. In (a) we plot the BIC of the learnt model on both the total dataset and on one sensitive subgroup. The absolute difference gives the BIC fairness in (b), which is normalised by the number of samples in the dataset in (c).

magnitude imbalancing, particularly with respect to the sensitive variable of gender, laid the groundwork for understanding the complexities involved in causal structure discovery and its susceptibility to biases.

The application of our methodologies to real-world datasets, such as the Adult dataset, provided valuable insights into the practical implications of our findings. This transition from controlled, artificial datasets to real-world data was crucial in validating our approaches in more complex, less predictable environments. Through this, we highlighted the importance of considering balanced representation in datasets to ensure the reliability and accuracy of causal discovery algorithms.

Further, we looked at the structural properties of graphs, examining how different configurations of chains, colliders, and forks, as well as the positioning of latent sensitive variables, influence the induced disparate impacts. This exploration not only contributed to a deeper understanding of the algorithmic behaviour in the presence of unobserved confounders but also underscored the challenges faced in real-world applications.

In summary, the methodologies delineated in this chapter form the backbone of our research, providing robust frameworks for examining and addressing fairness and bias in ML. The findings and insights gained from these methodological explorations set the stage for the subsequent chapters, where we will discuss their implications, and suggest potential pathways for future research in this vital field.

Throughout this chapter, our methodologies and results have collectively addressed our foundational research questions. The integration of causal inference in MARL (Section 5.1) has provided approaches and theoretical insights, contributing to RQ1 and RQ2. Simultaneously, our examination of causality and disparate impacts of sensitive subgroups (Section 5.2) has shed light on aspects of fairness and bias in (causal) ML, addressing RQ3. These investigations demonstrate the broad applicability and impact of causal inference in both enhancing RL methods and ensuring ethical considerations in ML algorithms.

Chapter 6

Discussion

This chapter engages in a detailed examination of the interplay between MARL and causal inference, focusing on the potential biases inherent in causal discovery algorithms, particularly in the context of fairness and bias implications. MARL involves multiple agents that learn and adapt within a shared environment, which characterises many of the complexities in real-world systems. The employment of Dec-POMDPs for modelling MARL problems offers a structured framework for understanding agent interactions and decision-making processes under uncertainty.

Central to this discussion is the concept of 'action as intervention', a perspective that naturally merges methodologies from causal inference and MARL, and is a key framing of the argument throughout this thesis. This fusion presents both promising opportunities and significant challenges for future research. The integration of causal inference within MARL, for instance, paves the way for advanced learning strategies, data fusion, and the exploration of counterfactual scenarios. Conversely, scrutinising biases from a causal perspective directs attention towards the imperative of fairness in ML systems. As discussed, one advantage of this approach is that causal models are easily and intuitively understood by humans, and can thus be refuted given domain expertise, experiments, and even policy decisions. That said, this approach also means that the causal models themselves must be known, and a danger of learning incorrect and unfair models appears. Though we argue this is often a good trade-off, much research remains to be done. This resonates with global demands for research into responsible and transparent AI technologies.

The subsequent sections of this chapter are dedicated to two main themes. Firstly, the integration of causal frameworks within MARL is discussed, elucidating the enhancements and benefits they introduce, as detailed in Sections 5.1 and 6.3. Secondly, the analysis shifts to examining biases in causal discovery algorithms, assessing their differential impacts on various subgroups. This examination is aimed at contributing to the overarching objective of achieving fairness in ML systems, as elaborated in Section 5.2.

This investigation not only endeavours to bridge existing knowledge gaps but also underscores the necessity for further research into validation mechanisms and the development of robust tools for assessing the fairness of causal models. This thesis also sets a foundation for an in-depth exploration of the limitations and ethical considerations for the deployment of (causal) RL methods.

In the following sections, the narrative aims to tackle technical challenges while simultaneously engaging with broader academic dialogues in the pertinent fields. A particular emphasis is placed on the responsible and fair deployment of ML models.

6.1 Dissecting Key Trends

Causal Inference in RL: The combination of causal inference with RL, particularly through graphical models, has been shown to improve algorithm performance in various areas. This approach helps RL models better understand and interact with their environments. By identifying the causal relationships in these environments, we can develop RL algorithms that are not only smarter but also more universally applicable. This trend directly addresses **RQ2** and **RQ3**, demonstrating how causal models can make RL more efficient and sensitive to different groups in learning tasks.

Counterfactual Reasoning: Introducing counterfactual reasoning to RL has been effective in improving how algorithms learn from past actions, both in real-time (online) and from pre-existing data (offline). This approach allows for better evaluation of different strategies and a deeper understanding of the consequences of actions, which can lead to better decision-making in RL systems.

Deep and Model-Based RL: Combining deep learning with model-based RL has opened up new ways to deal with problems involving many variables. This combination takes advantage of deep learning's ability to represent multifaceted data and the strength of causal models in explaining and predicting outcomes, which is particularly useful in tackling real-world challenges.

Multi-agent RL in Cooperative Environments: Applying causal insights to multi-agent RL in cooperative settings seems to improve communication and teamwork among agents. This approach could lead to better collective performance, helping groups of agents to work together more effectively in changing environments.

6.2 Navigating Emerging Trends and Innovative Avenues

During the initial stages of this research, several new concepts emerged that are worth further exploration:

Learning Causal Paths: Researching how RL agents navigate environments with changing conditions, like fluctuating mazes, is a promising area. Adding complexities such as changing weather or the presence of other agents could provide deeper insights into how RL agents learn and adapt in dynamic situations.

Learning from Peers: This idea looks at how agents in a multi-agent RL setting can learn from each other through imitation, sharing knowledge, and specialisation. Understanding the best ways to facilitate this peer learning could be key to improving how groups of agents work together.

Partial Model Learning and Merging: In multi-agent scenarios, developing partial models that are causally accurate could help integrate knowledge from different agents without causing conflicts. Researching how to effectively create and merge these models could be valuable for advancing multi-agent RL.

These topics, along with others like the impact of incorrect causal models on RL performance, present an interesting mix of theoretical challenges and practical applications. A focused investigation into these areas could reveal new insights and deepen our understanding of the relationship between causality and RL. This research also paves the way for collaboration across different fields, promoting a comprehensive approach to address the complexities in this emerging area.

6.3 Integrating Causal Inference in (MA)RL

Our exploration into incorporating causal inference within MARL, as detailed in Section 5.1, aims to address **RQ2** and **RQ3**. This investigation has identified significant opportunities to enhance our understanding and capabilities of multi-agent systems. Integrating causal principles into MARL offers substantial improvements over traditional methods in several key areas: data fusion, off-policy and offline learning, counterfactual reasoning, and causal learning.

6.3.1 Relevance and Implications of Data Fusion

Data fusion involves merging datasets from different conditions and policies. This approach is crucial for enhancing the scope and accuracy of the models learnt. Identifying causal relationships among variables allows for a more systematic and effective integration of these datasets, enriching the learning environment for agents. This advancement is particularly significant for real-world applications where data often originate from diverse sources and under varied conditions.

6.3.2 Off-Policy and Offline Learning Enhancement

Causal inference can significantly improve off-policy and offline learning strategies. These methods are essential when resources are limited or online learning environments are unavailable. By applying a causal perspective, we can address biases in learning from pre-existing datasets, enhancing the reliability and effectiveness of these learning strategies. This approach marks a step toward more scalable and resource-efficient learning models.

6.3.3 Counterfactual Reasoning in Action Evaluation

Counterfactual reasoning allows for the assessment of alternate scenarios, which is invaluable in situations where taking actions could be costly. This capability improves the data-efficiency of learning algorithms, enabling the evaluation of hypothetical scenarios. In sequential decision-making tasks, the ability to simulate and assess alternative outcomes is crucial for effective policy development and decision-making.

6.3.4 Causal Learning: Bridging The Known and The Unknown

Causal learning focuses on identifying causal structures from observational data. When combined with model-based RL, it becomes a potent tool for model refinement and hypothesis testing. This combination enables agents to adapt to changing environments, potentially reducing the learning curve and improving model robustness.

6.3.5 Dec-POMDPs as Multi-Agent Causal Models

Transforming Dec-POMDPs into MACMs provides a structured way to model interactions and shared elements in multi-agent systems. MACMs offer a comprehensive framework to capture the complexities of multi-agent dynamics, especially in scenarios that differ from standard Dec-POMDP formulations. This approach allows for a more detailed analysis and understanding of the intricate dynamics in multi-agent environments.

6.3.6 Future Horizons

This exploration represents just the beginning of an exciting intersection between causal inference and RL. Our findings underscore the need for ongoing research into causal frameworks to enhance our understanding and application of multi-agent systems. As MARL continues to evolve to address increasingly complex and realistic scenarios, the integration of causal inference stands as a potential key to unlocking more efficient, robust, and interpretable MARL frameworks. A detailed examination of the limitations, backed by empirical evidence, is crucial for charting a clear course for future developments in this field.

Future research should focus on delving deeper into causal frameworks to provide a richer understanding that responds to **RQ2** and **RQ3** more comprehensively. Empirical studies will be essential to validate the theoretical advancements and to investigate their capacity to enhance aspects like sample efficiency, coordination among agents, and the impact on diverse groups. This approach not only opens up new avenues for advancing MARL but also contributes to the broader conversation on the responsible application of AI technologies, particularly in the context of fairness and equity.

As we have delved into the integration of causal frameworks within MARL, we've highlighted their potential to improve our understanding and application of these complex systems. The incorporation of causal principles promises enhancements in algorithmic performance, data fusion, and decision-making under uncertainty. This advancement is important, especially in the context of creating more efficient, robust, and interpretable MARL frameworks, tailored for real-world applications.

However, with these advancements comes a responsibility to scrutinise the methodologies we employ, particularly in terms of their fairness and bias implications. It is also important to acknowledge that the tools and algorithms we develop are not isolated from the societal and ethical dimensions in which they operate. Thus we must also turn our attention to a critical and often challenging aspect of AI and machine learning: the potential biases inherent in the causal discovery algorithms themselves.

Therefore, the subsequent section discusses our investigation into bias and disparate impacts that arise when applying causal discovery algorithms, particularly in sensitive applications.

6.4 Investigating Causal Disparities and Bias

Our research, as outlined in Section 5.2, investigates the potential negative impacts that can arise from using learnt causal graphs in fairness and policy decision-making. This analysis emphasises the need for more fair and robust causal discovery methods, especially when dealing with sensitive causal pathways and latent variables that may affect different subgroups differently [178, 179].

The experiments conducted, particularly those focusing on gender imbalances, illustrate the effects of class and magnitude imbalances on causal structure discovery. By manipulating sensitive variable representation in datasets (gender and race in our experiments), we have shown how biases in causal learning can manifest and potentially lead to unfair outcomes. This is especially poignant in real-world scenarios, as demonstrated in our analysis using the Adult dataset, where gender as a latent variable significantly influenced the discovery process. Furthermore, our exploration of structural properties in graphs revealed how different configurations, such as chains, colliders, and forks, affect the fairness metrics, underscoring the complexity of achieving fairness in ML algorithms.

The findings are significant for a wide range of stakeholders, including researchers, practitioners, policymakers, and affected communities. As ML algorithms increasingly influence decision-making, it is crucial to promote fairness to avoid perpetuating or creating biases. Our study underscores the perils inherent in causal discovery algorithms that neglect subgroup distinctions, emphasising the imperative to rectify these biases for the development of equitable and unbiased ML systems.

An important next step is to develop and implement validation mechanisms, such as causal fairness metrics, to test the robustness of causal models. These metrics could help evaluate and guide causal models toward fairer outcomes. However, challenges arise, such as issues related to Pareto optimality [180], which suggests that achieving fairness across all definitions might be unattainable [181, 182]. This raises a critical question: should definitions of fairness be informed by judicial and regulatory frameworks, providing a basis for fairness in causal inference [183]?

Our case studies motivate a need for engagement in the academic community - we need to assess the real-world impact of applying methods that aim to replace domain knowledge. It is important to understand the extent of bias and whether causal models can mitigate or exacerbate it. Comparing past and current research could provide valuable insights, especially as causal fairness criteria become more prominent in academia.

Additionally, our research motivates the development of tools for assessing the fairness of learnt causal models. We identified a gap in the current framework for such evaluations. A step towards addressing this could be creating a suite of causal DAGs that include sensitive and latent variables. This suite could complement existing tools like Microsoft's CSuite [184], enhancing the evaluation of causal DAGs.

In conclusion, our investigation into biases in causal discovery algorithms and their impact on subgroup fairness propels us towards embracing the broader principles of fairness, accountability, and transparency in ML. By deepening our understanding of these challenges, we progress towards more equitable and inclusive AI systems, fostering social advancement and ensuring the responsible use of ML.

6.5 Conclusion

This discussion chapter has extensively considered the intersection of causal inference with MARL and investigated the inherent biases in causal discovery algorithms, addressing **RQ2** and **RQ3**. Through this exploration, we have highlighted significant advancements and potential future directions in MARL, counterfactual reasoning, and fairness in ML.

We began by integrating causal inference in MARL, uncovering its potential to enhance algorithmic performance, foster data fusion, and improve off-policy and offline learning strategies. This integration promises more efficient, robust, and interpretable MARL frameworks, opening new avenues for real-world applicability. The exploration of counterfactual reasoning and its impact on action evaluation further emphasises the evolution of RL, offering pathways for more effective policy development and decision-making in complex environments.

The investigation into biases in causal discovery algorithms, particularly concerning subgroup disparities, has shed light on the critical need for fairness in (causal) ML. We underscored the importance of developing validation mechanisms and fairness metrics to evaluate causal models, highlighting the challenges and complexities in achieving universally accepted definitions of fairness. This segment of the discussion stresses the vital role of responsible AI and the implications of ML decisions on broader societal contexts.

Future research directions have been identified, emphasising the necessity of empirical validation, deeper exploration of causal frameworks, and the development of new tools and methods to assess fairness in causal models. Our study sets the groundwork for a multidisciplinary approach, inviting collaboration across various fields to address the challenges and leverage the opportunities within this emerging research frontier.

In conclusion, this chapter contributes to the academic discourse by not only addressing technical challenges within the domains of MARL and causal inference but also by connecting used and the limitations of this study would provide a comprehensive closure to this section.

Chapter 7

Conclusions

Based on the discussion in Section 1.1, this thesis undertook an exploration at the intersection of RL and causality, focusing mainly on identifying trends rather than conducting experiments. Here are the responses to the proposed research questions based on the findings:

RQ1. Analysis of existing RL methods showed a variation in causal understanding across different RL and MARL algorithms. It was observed that classical algorithms largely lack the theoretical ability to exhibit performance expected of a true counterfactual reasoning agent, especially in low data resource scenarios. This affirms the initial hypothesis regarding the limitations in causal understanding in traditional RL methods.

RQ2. Introducing a causal model was found to improve the sample efficiency and coordination of learning agents in the single agent domain. By comparing RL environments with and without causal models/methods, it was clear that causal models enhance the rate of skill acquisition and learning effectiveness. This finding supports the hypothesis that causal methods can bolster the reasoning capability and sample efficiency of learning agents. Furthermore, It was found that there is not enough existing research progress to determine the contribution causal approaches can offer to decentralised learning tasks, especially with regard to multi-agent RL.

RQ3. Investigation revealed that causal learning can indeed lead to disparate impacts on sensitive subgroups within learning tasks. Applying learnt causal models to decision tasks showed evidence of disparate impacts and biased outcomes, corroborating the hypothesis that the application of causal learning could lead to biased outcomes affecting certain subgroups adversely.

The broader analysis in this study emphasises the potential benefits and challenges of integrating causal methods within RL and multi-agent RL scenarios. The intersections between causality and RL, although under-explored, present a viable path for improving the effectiveness and fairness of learning algorithms. This thesis lays a groundwork for further exploration into this interdisciplinary domain, demonstrating the potential for enhanced reasoning, efficiency, and nuanced understanding of learning paradigms through the fusion of causal inference and RL.

Moving forward, there are several directions for future work and exploration based on the findings of this thesis. Here are some potential next steps:

1. **Experimental Validation:** While this study primarily focused on identifying trends, there is a clear need for more experimental work to validate the findings. Future work could include designing and conducting experiments to confirm the effectiveness and efficiency gains from integrating causal models in RL and MARL environments.

2. **Decentralised Learning Tasks:** The impact of causal methods on decentralised learning tasks, especially within multi-agent RL, remains largely unexplored. Future research could consider experimenting with causal approaches in decentralised learning scenarios to better understand the benefits or challenges posed.

3. **Bias Mitigation:** Given the potential for causal learning to induce disparate impacts on sensitive subgroups, developing strategies to mitigate such biases is crucial. Future work could investigate methods to detect and correct biases arising from causal learning applications in decision tasks.

4. **Algorithm Development:** Creating new algorithmic frameworks that blend causal reasoning with RL could be a significant step towards addressing the research questions posed in this thesis more robustly. Developing algorithms that can naturally incorporate causal inference could significantly advance the field.

5. **Benchmarking and Metrics:** Establishing benchmarks and metrics to evaluate the causal understanding and effectiveness of RL and MARL algorithms could lead to a more systematic assessment and comparison. Future work could involve creating standardised benchmarks and evaluation protocols to facilitate a deeper understanding of the interplay between causality and RL.

The possibility of merging causal inference with RL opens up a wide range of research opportunities. The initial exploration conducted in this thesis lays the groundwork for further investigations that could potentially provide more nuanced insights and robust solutions at the intersection of causality and learning paradigms.

Bibliography

[1] David Hume. *An enquiry concerning human understanding.* Hackett, 2 edition, 1993.

[2] St John M M Grimbly, Jonathan Shock, and Arnu Pretorius. Causal multi-agent reinforcement learning: Review and open problems, 2021.

[3] St John M M Grimbly. Climbing the ladder: A survey of counterfactual methods in decision making processes. 2020.

[4] St John M M Grimbly. World models and predictive methods in deep reinforcement learning: A survey. 2020.

[5] Arnu Pretorius, Kale ab Tessera, Andries P. Smit, Claude Formanek, St John Grimbly, Kevin Eloff, Siphelele Danisa, Lawrence Francis, Jonathan Shock, Herman Kamper, Willie Brink, Herman Engelbrecht, Alexandre Laterre, and Karim Beguir. Mava: a research framework for distributed multi-agent reinforcement learning. 2021.

[6] Judea Pearl. *Causality.* Cambridge university press, 2009.

[7] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.

[8] Daniel Kahneman. *Thinking, fast and slow.* Macmillan, 2011.

[9] Karl Popper. *The logic of scientific discovery.* Routledge, 2005.

[10] Sandra Harding. *Can theories be refuted?: Essays on the Duhem-Quine thesis*, volume 81. Springer Science & Business Media, 1975.

[11] Ronald A. Fisher. Cancer and smoking. *Nature*, 182:596–596, 1958.

[12] Sir Ronald A. Fisher. Smoking: The cancer controversy. 1960.

[13] L. Penrose. Cancer and smoking. *Nature*, 182:1178–1178, 1958.

[14] Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. *Fundamentals of clinical trials.* Springer, 2015.

[15] Judea Pearl. *Causality.* Cambridge university press, 2009.

[16] Judea Pearl and Dana Mackenzie. *The Book of Why.* Basic Books, New York, 2018. ISBN 978-0-465-09760-9.

[17] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, February 2019. ISSN 0001-0782. doi: 10.1145/3241036. URL https://doi.org/10.1145/3241036.

[18] Mauricio Gonzalez-Soto and Felipe Orihuela Espina. Reinforcement learning is not a causal problem. *arXiv preprint arXiv:1908.07617*, 2019.

[19] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. unpublished, 2020.

[20] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[21] Judea Pearl and Azaria Paz. *Graphoids: A graph-based logic for reasoning about relevance relations.* University of California (Los Angeles). Computer Science Department, 1985.

[22] Sewall Wright. Systems of mating. i. the biometric relations between parent and offspring. *Genetics*, 6(2):111, 1921.

[23] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, Cambridge, MA, USA, 2017.

[24] Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.

[25] Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. 2005.

[26] James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

[27] Ben Glocker, Mirco Musolesi, Jonathan Richens, and Caroline Uhler. Causality in digital medicine. *Nature Communications*, 12(1), 2021.

[28] P. Spirtes, C. Glymour, and R. Scheines. Causation, prediction, and search, second edition. In *Adaptive computation and machine learning*, 2000.

[29] Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, 8 (3):255–268, 1991.

[30] Christopher Meek. *Graphical models: selecting causal and statistical models.* Phd thesis, Carnegie Mellon University, 1997.

[31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[32] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[33] Diviyan Kalainathan. *Generative Neural Networks to infer Causal Mechanisms : algorithms and applications*. Theses, Université Paris Saclay (COmUE), December 2019. URL https://theses.hal.science/tel-02528204.

[34] Alan M Turing. Lecture to the london mathematical society on 20 february 1947. *MD COMPUTING*, 12:390–390, 1995.

[35] Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice*, pages 91–121. Springer, 1984.

[36] A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950. ISSN 1460-2113. doi: 10.1093/mind/LIX.236.433.

[37] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

[38] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[39] Karl Tuyls, Julien Pérolat, Marc Lanctot, Joel Z. Leibo, and Thore Graepel. A generalised method for empirical game theoretic analysis. In *AAMAS*, 2018.

[40] Massimo Silvetti and Tom Verguts. Reinforcement learning, high-level cognition, and the human brain. 2012.

[41] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[42] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2019.

[43] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017.

[44] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments, 2020.

[45] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation, 2019.

[46] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

[47] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[48] Franccois Chollet. On the measure of intelligence. *ArXiv*, abs/1911.01547, 2019.

[49] Marvin Lee Minsky, editor. *Semantic Information Processing*. MIT Press, 1968.

[50] David Hume. *A Treatise of Human Nature*. Clarendon Press, 1739.

[51] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[52] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[53] Jürgen Schmidhuber. Annotated history of modern ai and deep learning. *arXiv preprint arXiv:2212.11279*, 2022.

[54] S-I Amari. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on computers*, 100(11):1197–1206, 1972.

[55] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[56] Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Conference on learning theory*, pages 2306–2327. PMLR, 2020.

[57] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

[58] Fabian Bernd Fuchs and Petar Veličković. Universality of neural networks on sets and graphs. In *The Second Blogpost Track at ICLR 2023*, 2023. URL https://openreview.net/forum?id=ynwoL965IM.

[59] Yann LeCun, D Touresky, G Hinton, and T Sejnowski. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pages 21–28. San Mateo, CA, USA, 1988.

[60] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[61] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.

[62] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[63] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[64] Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, second edition edition, 2018.

[65] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[66] Maor Gaon and Ronen I. Brafman. Reinforcement learning with non-markovian rewards. In *AAAI*, 2020.

[67] Steven D Whitehead and Long-Ji Lin. Reinforcement learning of non-markov decision processes. *Artificial intelligence*, 73(1-2):271–306, 1995.

[68] Sultan Javed Majeed and Marcus Hutter. On q-learning convergence for non-markov decision processes. In *IJCAI*, volume 18, pages 2546–2552, 2018.

[69] David L Poole and Alan K Mackworth. *Artificial Intelligence: foundations of computational agents*. Cambridge University Press, 2010.

[70] K. Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1964.

[71] Michael Negnevitsky. *Artificial intelligence: a guide to intelligent systems*. Pearson education, 2005.

[72] Claude E Shannon. Xxii. programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(314): 256–275, 1950.

[73] Garry Kasparov. *Deep thinking: where machine intelligence ends and human creativity begins*. Hachette UK, 2017.

[74] Richard E Bellman et al. Dynamic programming, ser. *Cambridge Studies in Speech Science and Communication. Princeton University Press, Princeton*, 1957.

[75] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[76] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

[77] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

[78] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[79] Olivier Cappé , Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3), jun 2013. doi: 10.1214/13-aos1119. URL https://doi.org/10.1214%2F13-aos1119.

[80] Anusha Nagabandi, Kurt Konoglie, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation, 2019.

[81] Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *Intelligencesigart Bulletin*, July 1991.

[82] Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration, 2016.

[83] Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I. Jordan, Joseph E. Gonzalez, and Sergey Levine. Model-based value estimation for efficient model-free reinforcement learning, 2018.

[84] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization, 2019.

[85] David Ha and Jürgen Schmidhuber. World models. 2018. doi: 10.5281/ZENODO.1207631. URL https://zenodo.org/record/1207631.

[86] Martin Dresler, Stefan P. Koch, Renate Wehrle, Victor I. Spoormaker, Florian Holsboer, Axel Steiger, Philipp G. Sämann, Hellmuth Obrig, and Michael Czisch. Dreamed movement elicits activation in the sensorimotor cortex. *Current Biology*, 21:1833–1837, 2011.

[87] Sarah Fiona Schoch, Maren Jasmin Cordi, Michael Schredl, and Björn Rasch. The effect of dream report collection and dream incorporation on memory consolidation during sleep. *Journal of Sleep Research*, 28, 2019.

[88] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *ArXiv*, abs/1912.01603, 2020.

[89] Michael Wooldridge. *An introduction to multiagent systems*. John wiley & sons, 2009.

[90] John Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[91] W. Deming, J. Neumann, and O. Morgenstern. Theory of games and economic behavior. *Journal of the American Statistical Association*, 40:263, 1944.

[92] François Bousquet and Christophe Le Page. Multi-agent simulations and ecosystem management: a review. *Ecological modelling*, 176(3-4):313–332, 2004.

[93] Cars H Hommes. Heterogeneous agent models in economics and finance. *Handbook of computational economics*, 2:1109–1186, 2006.

[94] Jacques Ferber and Gerhard Weiss. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-wesley Reading, 1999.

[95] M. Littman. Markov games as a framework for multi-agent reinforcement learning. 1994.

[96] Yoav Shoham, Rob Powers, and Trond Grenager. Multi-agent reinforcement learning: a critical survey. 2003.

[97] Sven Gronauer and K. Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, pages 1–49, 2021.

[98] S. Meganck, S. Maes, B. Manderick, and P. Leray. Distributed learning of multi-agent causal models. pages 285–288, 2005. doi: 10.1109/IAT.2005.66.

[99] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9):3826–3839, 2020.

[100] Roberta Raileanu, Emily L. Denton, Arthur D. Szlam, and R. Fergus. Modeling others using oneself in multi-agent reinforcement learning. 2018.

[101] P. Sharma, Rolando Fernandez, Erin G. Zaroukian, M. Dorothy, Anjon Basak, and Derrik E. Asher. Survey of recent multi-agent reinforcement learning algorithms utilizing centralized training. 2021.

[102] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. 2021.

[103] Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz De Cote. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2017.

[104] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa. Robocup: The robot world cup initiative. In *Proceedings of the first international conference on Autonomous agents*, pages 340–347, 1997.

[105] E. Hansen, D. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. 2004.

[106] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.

[107] Elias Bareinboim. Causal reinforcement learning. ICML 2020, 2020. URL https://icml.cc/virtual/2020/tutorial/5752.

[108] Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

[109] Andrew Forney, J. Pearl, and Elias Bareinboim. Counterfactual data-fusion for online reinforcement learners. 2017.

[110] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

[111] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[112] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[113] Erica EM Moodie, Bibhas Chakraborty, and Michael S Kramer. Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics*, 40(4):629–645, 2012.

[114] Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, and Susan A Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225, 2014.

[115] L. Wang, A. Rotnitzky, Xihong Lin, R. Millikan, and P. Thall. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association*, 107:493 – 508, 2012.

[116] Peter L Spirtes, Christopher Meek, and Thomas S Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

[117] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[118] David A Stephens, Widemberg S Nobre, Erica EM Moodie, and Alexandra M Schmidt. Causal inference under mis-specification: adjustment based on the propensity score. *arXiv preprint arXiv:2201.12831*, 2022.

[119] P. Spirtes, C. Glymour, and R. Scheines. Causation, prediction, and search, second edition. 2000.

[120] Samuel J Gershman. Reinforcement learning and causal models. *The Oxford handbook of causal reasoning*, 1:295, 2017.

[121] A Philip Dawid. Conditional independence for statistical operations. *The Annals of Statistics*, 8(3):598–617, 1980.

[122] Matthias Brand, Kimberly S Young, Christian Laier, Klaus Wölfling, and Marc N Potenza. Integrating psychological and neurobiological considerations regarding the development and maintenance of specific internet-use disorders: An interaction of person-affect-cognition-execution (i-pace) model. *Neuroscience & Biobehavioral Reviews*, 71:252–266, 2016.

[123] Junchi Liang and Abdeslam Boularias. Inferring time-delayed causal relations in pomdps from the principle of independence of cause and mechanism. volume 2, page 1944–1950, Aug 2021. doi: 10.24963/ijcai.2021/268. URL https://www.ijcai.org/proceedings/2021/268.

[124] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. (arXiv:2005.01643), Nov 2020. URL http://arxiv.org/abs/2005.01643. arXiv:2005.01643 [cs, stat].

[125] Ann L Brown and Mary Jo Kane. Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive psychology*, 20(4):493–523, 1988.

[126] Ling Shao, Fan Zhu, and Xuelong Li. Transfer learning for visual categorization: A survey. *IEEE transactions on neural networks and learning systems*, 26(5): 1019–1034, 2014.

[127] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[128] Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandits: A causal approach. pages 1340–1346, 2017. doi: 10.24963/ijcai.2017/186. URL https://doi.org/10.24963/ijcai.2017/186.

[129] J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *NIPS 2007*, 2007.

[130] Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. 2019.

[131] Elias Bareinboim, Andrew Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. 2015.

[132] Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? *Advances in Neural Information Processing Systems 31*, 31, 2018.

[133] Neal J Roese. Counterfactual thinking. *Psychological bulletin*, 121(1):133, 1997.

[134] Kai Epstude and Neal J Roese. The functional theory of counterfactual thinking. *Personality and social psychology review*, 12(2):168–192, 2008.

[135] Antti Revonsuo. The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and brain sciences*, 23(6):877–901, 2000.

[136] Robert Stickgold, J Allen Hobson, Roar Fosse, and Magdalena Fosse. Sleep, learning, and dreams: off-line memory reprocessing. *Science*, 294(5544):1052–1057, 2001.

[137] P. Auer, Nicolò Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47:235–256, 2004.

[138] Andrew Forney and Elias Bareinboim. Counterfactual randomization: Rescuing experimental studies from obscured confounding. 2019.

[139] Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. 2016.

[140] Elias Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345 – 7352, 2016.

[141] W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

[142] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015.

[143] Athanasios Vlontzos, Bernhard Kainz, and Ciarán M Gilligan-Lee. Estimating categorical counterfactuals via deep twin networks. *Nature Machine Intelligence*, 5(2):159–168, 2023.

[144] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4), Nov 2014. ISSN 0883-4237. doi: 10.1214/14-sts486. URL http://dx.doi.org/10.1214/14-STS486.

[145] Elias Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. 2014.

[146] S. Lee, Juan David Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. 2019.

[147] P. Abbeel and A. Ng. Apprenticeship learning via inverse reinforcement learning. 2004.

[148] Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. *Advances in neural information processing systems*, 33, 2020.

[149] M. Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. 2017.

[150] Hal Ashton. Causal campbell-goodhart's law and reinforcement learning, 2021.

[151] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021.

[152] C. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[153] John Hicks et al. *Causality in economics.* Australian National University Press, 1980.

[154] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*, 2019.

[155] J. Harsanyi. Games with incomplete information played by "bayesian" players, i-iii: Part i. the basic model&. *Manag. Sci.*, 14:159–182, 1967.

[156] Frans A Oliehoek, Matthijs TJ Spaan, Nikos Vlassis, and Shimon Whiteson. Exploiting locality of interaction in factored dec-pomdps. pages 517–524, 2008.

[157] Guillermo Vigueras and Juan A Botia. Tracking causality by visualization of multi-agent interactions using causality graphs. In *International Workshop on Programming Multi-Agent Systems*, pages 190–204. Springer, 2007.

[158] A. Marcellesi. External validity: Is there still a problem? *Philosophy of Science*, 82:1308 – 1317, 2015.

[159] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345 – 7352, 2016.

[160] Sascha O. Becker. Using instrumental variables to establish causality. *The IZA World of Labor*, pages 250–250, 2016.

[161] C. Glymour, Kun Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

[162] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. *arXiv preprint arXiv:2006.10833*, 2020.

[163] Cong Su, Guoxian Yu, Jun Wang, Zhongmin Yan, and Lizhen Cui. A review of causality-based fairness machine learning. *Intelligence & Robotics*, 2(3):244–274, Aug 2022. doi: 10.20517/ir.2022.17.

[164] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. URL https://fairmlbook.org/.

[165] Rūta Binkytė-Sadauskienė, Karima Makhlouf, Carlos Pinzón, Sami Zhioua, and Catuscia Palamidessi. Causal discovery for fairness. (arXiv:2206.06685), 2008. URL http://arxiv.org/abs/2206.06685. arXiv:2206.06685 [cs, stat].

[166] Simon Caton and Christian Haas. Fairness in machine learning: A survey. (arXiv:2010.04053), Oct 2020. doi: 10.48550/arXiv.2010.04053. URL http://arxiv.org/abs/2010.04053. arXiv:2010.04053 [cs, stat].

[167] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Structural agnostic modeling: Adversarial learning of causal graphs. (arXiv:1803.04929), Jul 2022. doi: 10.48550/arXiv.1803.04929. URL http://arxiv.org/abs/1803.04929. arXiv:1803.04929 [stat].

[168] C. Glymour, Kun Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

[169] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvarinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 2006.

[170] Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company, Oct 2020. ISBN 978-0-393-63583-6.

[171] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks., May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[172] Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. (arXiv:1903.02278), Mar 2019. doi: 10.48550/arXiv.1903.02278. URL http://arxiv.org/abs/1903.02278. arXiv:1903.02278 [stat].

[173] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[174] Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *arXiv preprint arXiv:2206.06821*, 2022.

[175] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[176] Isabelle Guyon, Constantin Aliferis, Greg Cooper, and André Elisseeff. Datasets of the causation and prediction challenge. 2008.

[177] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

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

[179] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

[180] Susan Wei and Marc Niethammer. The fairness-accuracy pareto front, 2021.

[181] Christian Haas. The price of fairness-a framework to explore trade-offs in algorithmic fairness. 2019.

[182] Jon Kleinberg. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, pages 40–40, 2018.

[183] Michael Butterworth. The ico and artificial intelligence: The role of fairness in the gdpr framework. *Computer Law & Security Review*, 34(2):257–268, 2018.

[184] Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Deep end-to-end causal inference, 2022.

[185] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[186] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2000.

[187] Alison Gopnik, Laura Schulz, and Laura Elizabeth Schulz. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press, 2007.

[188] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[189] Azaria Paz, Judea Pearl, and Shmuel Ur. A new characterization of graphs based on interception relations. *Journal of Graph Theory*, 22(2):125–136, 1996.

[190] Dan Geiger. *The non-axiomatizability of dependencies in directed acyclic graphs*. University of California (Los Angeles). Computer Science Department, 1987.

[191] Ankur Ankan and Johannes Textor. pgmpy: A python toolkit for bayesian networks, 2023.

[192] Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python, 2019.

[193] Judea Pearl. [bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

[194] J. Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27:226 – 284, 1998.

[195] D. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322 – 331, 2005.

[196] J. Peters, D. Janzing, and B. Schölkopf. Elements of causal inference: Foundations and learning algorithms. 2017.

[197] I. Shpitser and J. Tian. On identifying causal effects. 2010.

[198] Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *arXiv preprint arXiv:2103.04850*, 2021.

[199] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. 2008.

[200] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvarinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 2006.

[201] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[202] Bryan E. Shepherd, Ryan T. Jarrett, and Lingjun Fu. Guido imbens, donald rubin, causal inference for statistics, social, and biomedical sciences: An introduction. new york: Cambridge university press. *Biometrics*, 72 4:1387–1388, 2016. URL https://api.semanticscholar.org/CorpusID:207068138.

[203] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition, 2008.

[204] Dominik Janzing, Rafael Chaves, and Bernhard Schölkopf. Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New Journal of Physics*, 18(9):093052, sep 2016. doi: 10.1088/1367-2630/18/9/093052. URL https://doi.org/10.1088%2F1367-2630%2F18%2F9%2F093052.

[205] Jonas Peters and Peter Bühlmann. Structural intervention distance (sid) for evaluating causal graphs. (arXiv:1306.1043), Apr 2014. URL http://arxiv.org/abs/1306.1043. arXiv:1306.1043 [stat].

[206] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, and Huawen Liu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(5):1483–1495, Sep 2019. ISSN 1545-5963, 1557-9964, 2374-0043. doi: 10.1109/TCBB.2016.2591526. arXiv:1502.02454 [cs].

[207] Yangyi Lu, A. Meisami, Ambuj Tewari, and Zhenyu Yan. Regret analysis of bandit problems with causal background knowledge. 2020.

[208] Jin Li, Ye Luo, and Xiaowei Zhang. Causal reinforcement learning: An instrumental variable approach. *Available at SSRN 3792824*, 2021.

[209] Maxime Gasse, Damien Grasset, Guillaume Gaudron, and Pierre-Yves Oudeyer. Causal reinforcement learning using observational and interventional data. *arXiv preprint arXiv:2106.14421*, 2021.

[210] A. Ghassami, Saber Salehkaleybar, N. Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. 2018.

[211] Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. 2019.

[212] M. Kocaoglu, A. Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. 2019.

[213] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33, 2020.

[214] J. Zhang and Elias Bareinboim. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. 2020.

[215] Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *arXiv preprint arXiv:2003.05623*, 2020.

[216] J. Zhang and Elias Bareinboim. Bounding causal effects on continuous outcomes. 2020.

[217] Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in neural information processing systems*, 33, 2020.

[218] J. Zhang and Elias Bareinboim. Can humans be out of the loop? 2020.

[219] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal markov decision processes: Learning good interventions efficiently. *arXiv preprint arXiv:2102.07663*, 2021.

[220] Tomer Galanti, Ofir Nabati, and Lior Wolf. A critical view of the structural causal model. *arXiv preprint arXiv:2002.10007*, 2020.

[221] Jonathan G Richens, Ciarán M. Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11, 2020.

[222] Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. *arXiv preprint arXiv:1304.7920*, 2013.

[223] Stephan Bongers and Joris M Mooij. From random differential equations to structural causal models: The stochastic case. *arXiv preprint arXiv:1803.08784*, 2018.

[224] J. Pearl. The causal foundations of structural equation modeling. 2012.

[225] R. MacLehose, S. Kaufman, J. Kaufman, and C. Poole. Bounding causal effects under uncontrolled confounding using counterfactuals. *Epidemiology*, 16:548–555, 2005.

[226] P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[227] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. Model-based reinforcement learning: A survey. *arXiv preprint arXiv:2006.16712*, 2020.

[228] Petter N. Kolm and Gordon Ritter. Modern perspectives on reinforcement learning in finance. *Econometrics: Mathematical Methods & Programming eJournal*, 2019.

[229] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[230] Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, S. Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking model-based reinforcement learning. *ArXiv*, abs/1907.02057, 2019.

[231] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[232] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[233] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application, 2018.

[234] Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. A study of reinforcement learning for neural machine translation, 2018.

[235] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language, 2017.

[236] Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

[237] Vojtěch Kovařík, Martin Schmid, Neil Burch, Michael Bowling, and Viliam Lisỳ. Rethinking formal models of partially observable multiagent decision making. *arXiv preprint arXiv:1906.11110*, 2019.

[238] Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems*, 34(1), Dec 2019. ISSN 1573-7454. doi: 10.1007/s10458-019-09433-x. URL http://dx.doi.org/10.1007/s10458-019-09433-x.

[239] C. S. D. Witt, Bei Peng, Pierre-Alexandre Kamienny, Philip H. S. Torr, Wendelin Böhmer, and S. Whiteson. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *ArXiv*, abs/2003.06709, 2020.

[240] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. 2016.

[241] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010.

[242] Junqi Jin, C. Song, Han Li, Kun Gai, J. Wang, and Weinan Zhang. Real-time bidding with multi-agent reinforcement learning in display advertising. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.

[243] J. Lussange, I. Lazarevich, S. Bourgeois-Gironde, Stefano Palminteri, and B. Gutkin. Modelling stock markets by multi-agent reinforcement learning. *Computational Economics*, 57:113–147, 2020.

[244] Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *arXiv preprint arXiv:2106.03400*, 2021.

[245] Mikayel Samvelyan, Tabish Rashid, C. S. D. Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob N. Foerster, and S. Whiteson. The starcraft multi-agent challenge. 2019.

[246] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. pages 4295–4304, 2018.

[247] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. 32(1), 2018.

[248] Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. Roma: Multi-agent reinforcement learning with emergent roles. *arXiv preprint arXiv:2003.08039*, 2020.

[249] Oriol Vinyals, I. Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, J. Chung, David H. Choi, Richard Powell, Timo Ewalds, P. Georgiev, Junhyuk Oh, Dan Horgan, M. Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, J. Agapiou, Max Jaderberg, A. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, D. Budden, Yury Sulsky, James Molloy, T. Paine, Caglar Gulcehre, Ziyu Wang, T. Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, pages 1–5, 2019.

[250] K. Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and T. Başar. Fully decentralized multi-agent reinforcement learning with networked agents. 2018.

[251] Virgile Landeiro and A. Culotta. Robust text classification under confounding shift. *J. Artif. Intell. Res.*, 63:391–419, 2018.

[252] Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Addressing distribution shift in online reinforcement learning with offline datasets. 2020.

[253] Russell Mendonca, Xinyang Geng, Chelsea Finn, and Sergey Levine. Meta-reinforcement learning robust to distributional shift via model identification and experience relabeling. *arXiv preprint arXiv:2006.07178*, 2020.

[254] Prabuchandran K.J., Hemanth Kumar A.N, and Shalabh Bhatnagar. Multi-agent reinforcement learning for traffic signal control. pages 2529–2534, 2014. doi: 10.1109/ITSC.2014.6958095.

[255] Woojun Kim, Jongeui Park, and Y. Sung. Communication in multi-agent reinforcement learning: Intention sharing. 2021.

[256] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, D. Strouse, Joel Z. Leibo, and N. D. Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. 2019.

[257] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. Tarmac: Targeted multi-agent communication. pages 1538–1546, 2019.

[258] Shayegan Omidshafiei, Dong-Ki Kim, Miao Liu, G. Tesauro, M. Riemer, Chris Amato, Murray Campbell, and J. How. Learning to teach in cooperative multiagent reinforcement learning. 2019.

[259] Matthieu Zimmer, Paolo Viappiani, and Paul Weng. Teacher-student framework: a reinforcement learning approach. 2014.

[260] Ercüment Ilhan, J. Gow, and Diego Perez Liebana. Teaching on a budget in multiagent deep reinforcement learning. *2019 IEEE Conference on Games (CoG)*, pages 1–8, 2019.

[261] Saeed Kaviani, Bo Ryu, Ejaz Ahmed, Kevin A Larson, Anh Le, Alex Yahja, and Jae H Kim. Robust and scalable routing with multi-agent deep reinforcement learning for manets. *arXiv preprint arXiv:2101.03273*, 2021.

[262] Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable multi-agent reinforcement learning for networked systems with average reward. *arXiv preprint arXiv:2006.06626*, 2020.

[263] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pages 5571–5580. PMLR, 2018.

[264] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[265] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[266] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

[267] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998(746-752):2, 1998.

[268] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

[269] David H Wolpert, Kagan Tumer, and Keith Swanson. Optimal wonderful life utility functions in multi-agent systems. 2000.

[270] Sam Devlin, Logan Yliniemi, Daniel Kudenko, and Kagan Tumer. Potential-based difference rewards for multiagent reinforcement learning. pages 165–172, 2014.

[271] Frans A Oliehoek and Christopher Amato. *A concise introduction to decentralized POMDPs*. Springer, 2016.

[272] M. Gonzalez-Soto, L. Sucar, and H. Escalante. Causal games and causal nash equilibrium. *Res. Comput. Sci.*, 149:123–133, 2020.

[273] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, D. Strouse, Joel Z. Leibo, and N. D. Freitas. Intrinsic social motivation via causal influence in multi-agent rl. *ArXiv*, abs/1810.08647, 2018.

[274] Simon Vanneste, Astrid Vanneste, Kevin Mets, Ali Anwar, Siegfried Mercelis, Steven Latré, and Peter Hellinckx. Learning to communicate using counterfactual reasoning. *arXiv preprint arXiv:2006.07200*, 2020.

[275] S. Maes, S. Meganck, and B. Manderick. Inference in multi-agent causal models. *Int. J. Approx. Reason.*, 46:274–299, 2007.

[276] Alfred J. Lotka. Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry*, 14(3):271–274, Mar 1910. ISSN 0092-7325. doi: 10.1021/j150111a004.

[277] NARENDRA S. GOEL, SAMARESH C. MAITRA, and ELLIOTT W. MONTROLL. On the volterra and other nonlinear models of interacting populations. *Reviews of Modern Physics*, 43(2):231–276, Apr 1971. doi: 10.1103/RevModPhys.43.231.

[278] User Ninjatacoshell. Kernel machine, 2016. Available on: https://commons.wikimedia.org/wiki/File:Kernel_Machine.png. Accessed: 27 July 2023. File: Kernel_Machine.png.

[279] Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, K. Selcuk Candan, and Huan Liu. Evaluation methods and measures for causal learning algorithms. (arXiv:2202.02896), Feb 2022. doi: 10.48550/arXiv.2202.02896. URL http://arxiv.org/abs/2202.02896. arXiv:2202.02896 [cs, stat].

[280] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. *Learning Functional Causal Models with Generative Neural Networks*. 2018. doi: 10.1007/978-3-319-98131-4. URL http://arxiv.org/abs/1709.05321. arXiv:1709.05321 [stat].

[281] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):115:1–115:35, Jul 2021. ISSN 0360-0300. doi: 10.1145/3457607.

[282] Luca Oneto and Silvia Chiappa. *Fairness in Machine Learning*, page 155–196. Studies in Computational Intelligence. Springer International Publishing, Cham, 2020. ISBN 978-3-030-43883-8. doi: 10.1007/978-3-030-43883-8_7. URL https://doi.org/10.1007/978-3-030-43883-8_7.

[283] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys*, 55(3):51:1–51:44, Feb 2022. ISSN 0360-0300. doi: 10.1145/3494672.

[284] Jonas Peters. Elements of causal inference: foundations and learning algorithms. 2017.

[285] Jonas Peters and Peter Bühlmann. Structural intervention distance (sid) for evaluating causal graphs. (arXiv:1306.1043), Apr 2014. URL http://arxiv.org/abs/1306.1043. arXiv:1306.1043 [stat].

[286] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, page 1–7, Gothenburg Sweden, May 2018. ACM. ISBN 978-1-4503-5746-3. doi: 10.1145/3194770.3194776. URL https://dl.acm.org/doi/10.1145/3194770.3194776.

[287] Silvia Chiappa. Path-specific counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(0101):7801–7808, Jul 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33017801.

[288] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer, New York, NY, 1993. ISBN 978-1-4612-7650-0. doi: 10.1007/978-1-4612-2748-9. URL http://link.springer.com/10.1007/978-1-4612-2748-9.

[289] Christopher Meek. *Graphical models: selecting causal and statistical models*. Phd thesis, Carnegie Mellon University, 1997.

[290] Hamed Nilforoshan, Johann Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. (arXiv:2207.05302), Jul 2022. doi: 10.48550/arXiv.2207.05302. URL http://arxiv.org/abs/2207.05302. arXiv:2207.05302 [cs].

[291] David Maxwell Chickering. Optimal structure identification with greedy search.

[292] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[293] Karl Pearson. On the theory of contingency and its relation to association and normal correlation. *(No Title)*, 1904.

[294] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[295] Wei-Yin Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014.

[296] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong but many are useful: variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv:1801.01489*, 1, 2018.

[297] Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67, 2011.

[298] David A Freedman. Statistical models and shoe leather. *Sociological methodology*, pages 291–313, 1991.

[299] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[300] Jean-Marie Dufour and Eric Renault. Short run and long run causality in time series: theory. *Econometrica*, pages 1099–1125, 1998.

[301] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85 (2):461, 2000.

[302] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press, 2009.

Appendix A

Causality Appendix

This appendix complements Chapter 2. The structure of this chapter follows the order in which references are made to this appendix from the main body of the thesis. The aim is to offer additional details, proofs, or algorithms that might not directly contribute to the main argument of the thesis, but could aid in understanding or intuition. Its purpose is to clearly demonstrate how causality differs from simple correlation and why this distinction is crucial in data interpretation. The examples at the end of the chapter are selected for their ability to clearly show instances where relying solely on statistical correlations can lead to incorrect conclusions.

A.1 Graphoids as Graph Foundations

Consider a three-part knowledge system where an agent must assess the truth of statement $I(x, z, y)$, meaning "statement x is independent of y given z." In other words, given knowledge of z, the state of y is not required to determine x. This introduces the idea of a *dependence model* M, which consists of a subset of triplets (X, Z, Y) where the statement "X is independent of Y given Z" holds.

Definition A.1.1 (Graphoid)**.** *One defines a semi-graphoid as a dependency model that is closed under axioms 1-4, while a* graphoid *is additionally closed under axiom 5 [21]:*

1. *Symmetry:* $I(X, Z, Y) \Leftrightarrow I(Y, Z, X)$
   Example: If knowing the brand of a car Z renders the relationship between its price X and fuel efficiency Y independent, then knowing the brand should also render the relationship between fuel efficiency and price independent.

2. *Decomposition:* $I(X, Z, Y \cup W) \Rightarrow I(X, Z, Y)$ & $I(X, Z, W)$
   Example: If the price of a car X is independent of the combination of its colour and weight $Y \cup W$ given the brand Z, then the price is also independent of its colour and weight separately given the brand.

3. *Weak Union:* $I(X, Z, Y \cup W) \Rightarrow I(X, Z \cup W, Y)$
   Example: If the price of a car X is independent of the combination of its colour and weight $Y \cup W$ given the brand Z, then the price is also independent of the colour given the brand and weight $Z \cup W$.

4. *Contraction:* $I(X, Z, Y)$ & $I(X, Z \cup Y, W) \Rightarrow I(X, Z, Y \cup W)$
   Example: If the fuel efficiency of a car X is independent of its colour Y given its brand Z, and is also independent of its weight W given the brand and colour $Z \cup Y$, then the fuel efficiency is independent of the combination of colour and weight given the brand.

5. *Intersection:* $I(X, Z \cup W, Y)$ & $I(X, Z \cup Y, W) \Rightarrow I(X, Z, Y \cup W)$
   Example: If the price of a car X is independent of its colour Y given the brand and weight $Z \cup W$, and is also independent of its weight W given the brand and colour $Z \cup Y$, then the price is independent of the combination of colour and weight given the brand.

These properties are well motivated by Pearl [15, pg. 12]. These graphoid properties are favourable in settings where reasoning conditionally is important. As such, graphoids can be used as a basis for defining graphs.

Definition A.1.2 (Graph)**.** *An (undirected) graph is a pair $G = (V, E)$ where V is a set of elements called vertices and E is a set of elements called edges. A directed graph is a graph in which the edges E have a topological ordering.*

If there exists an undirected graph G such that

$$I(X, Z, Y) \Leftrightarrow \langle X, Z, Y \rangle_G,$$

then the graphoid is said to be *graph-induced*. One can show that if, in addition to the graphoid axioms, strong union and transitivity hold, then the condition is both necessary and sufficient [189]. In other words, these axioms form a complete characterisation of undirected graphs.

A.1.1 DAG-induced Graphoids

A graphoid can be classified as *DAG-induced* when there is a directed acyclic graph (DAG) D satisfying the condition

$$I(X, Z, Y) \Leftrightarrow \langle X, Z, Y \rangle_D,$$

with $\langle X, Z, Y \rangle_D$ indicating d-separation in D. This essentially means that conditional independence can be inferred from the structure of a Bayesian network (refer to subsection 2.5). In this context, Geiger [190] demonstrates that no finite set of axioms completely characterises conditional independencies in a DAG.

Theorem A.1.1 (Probabilistic Implications of d-Separation [15])**.** *For a DAG G, if the sets X and Y are d-separated by Z, then in every distribution that aligns with G, X is independent of Y given Z. On the other hand, if X and Y are not d-separated by Z in G, then there exists at least one distribution consistent with G where X and Y exhibit dependency conditional on Z.*

Theorem A.1.2 (Observational Equivalence)**.** *Two DAGs are deemed* observationally equivalent *if they possess identical skeletons and v-structures. This means they share configurations like $X \rightarrow Y \leftarrow Z$, where X, Y, Z are variables within the models.*

The concept of observational equivalence is crucial as it delineates the boundaries within which causal directions can be inferred solely from (non-temporal) data and probabilities. To extend our understanding beyond this equivalence, approaches beyond mere observation—such as experimentation and intervention—are necessary. This topic is further explored in subsequent sections.

- Strong Union: $I(X, Z, Y) \implies I(X, Z \cup W, Y)$
  Example: Assume we have four variables: the price of a car X, its brand Z, its colour Y, and its weight W. If the price of a car is independent of its colour given its brand, then adding the weight of the car to the conditioning set should not affect this independence.
- Transitivity: $I(X, Z, Y) \implies (\forall\, \gamma \notin X \cup Y \cup Z,\quad I(X, Z, \gamma)$ or $I(\gamma, Z, Y))$
  Example: Continuing with the car analogy, assume we have a fifth variable: the engine size γ. If the price of the car X is independent of its colour Y given its brand Z, then either the price of the car is independent of its engine size given the brand, or the engine size is independent of the colour given the brand. This axiom essentially posits a sort of "transitive" independence amongst variables outside the given conditioning set.

Assuming that we remove all the arrowheads in a directed graph G, we are left with the *skeleton* of G. We define a path as any unbroken, non-intersecting route in the graph, regardless of direction of the edges involved. A directed path is a path in which every edge points in the same direction. That is, every edge in the path has an arrow point from the first to the second vertex.

Though directed graphs may include directed cycles, they may not include a self-contained edge. For example, $X \to X$ is not a well defined edge in this context, though there may exist a directed path $X \to Y \to X$.

A.2 Other Causal Frameworks

In the study of causality, while graphical models such as those proposed by Judea Pearl have been foundational, they are not the only paradigms in existence. This section ventures into alternative causal frameworks that offer different approaches and mathematical tools for causal inference. The aim is to explore these frameworks as complements (or sometimes alternatives) to graphical causal models. We begin with the potential outcomes framework, more commonly known as the Rubin Causal Model, which has gained significant traction particularly in economics and social sciences.

A.2.1 Potential Outcomes Framework

The potential outcomes framework [20, 201, 202], offers a complementary perspective to Judea Pearl's graphical approach in causal inference. It provides a mathematical framework for analysing causal effects and underpins numerous statistical methodologies. This framework quantifies causal effects by considering the potential outcomes of interventions. Here, the potential outcomes relate to a *unit*—the entity under causal investigation. For instance, in a study on student performance, the unit would be the individual student, and the potential outcomes might be their grades under various influencing factors (like family income or school type). It forms the basis for methods such as matching, instrumental variables, and propensity score analysis.

With a binary treatment variable T, the potential outcome for unit i under treatment ($T = 1$) is denoted as $Y_i(1)$, and under control ($T = 0$) as $Y_i(0)$. The observed outcome Y_i for unit i is then expressed as:

$$Y_i = T \cdot Y_i(1) + (1 - T) \cdot Y_i(0). \tag{A.1}$$

The causal effect for unit i is defined by the difference between these potential outcomes;

$$\tau_i = Y_i(1) - Y_i(0). \tag{A.2}$$

A.2.2 Assumptions

The potential outcomes framework rests on several assumptions:

- **Stable Unit Treatment Value Assumption (SUTVA):** This posits that the potential outcomes for any unit are not influenced by the treatments applied to other units.
- **Ignorability:** It assumes that the treatment assignment is independent of the potential outcomes, conditional on observed covariates.
- **Positivity:** It holds that every unit has a nonzero probability of receiving each level of treatment.

Example 16 (Simple Worked Example)**.** *Consider a simple example to demonstrate the potential outcomes framework: evaluating the effect of a new training program on employee productivity. Suppose a company introduces a training program to enhance productivity. The* potential outcomes *for an employee i are:*

$$\begin{aligned} Y_i(1) &: \textit{Productivity if employee } i \textit{ undergoes the training.} \\ Y_i(0) &: \textit{Productivity if employee } i \textit{ does not undergo the training.} \end{aligned}$$

The Causal Effect$_i$ *is then* $Y_i(1) - Y_i(0)$*. Assuming data from 10 employees, with 5 in the treatment group and 5 in the control group, their observed outcomes might be:*

$$\begin{aligned} \textit{Treatment Group:} \quad & Y = \{100, 110, 120, 90, 105\} \\ \textit{Control Group:} \quad & Y = \{95, 90, 80, 85, 92\} \end{aligned}$$

The challenge in causal inference is the inability to observe both potential outcomes for each employee. However, the average treatment effect (ATE) can be estimated from the data:

$$ATE = \frac{\sum Y_i(1)}{N_{Treatment}} - \frac{\sum Y_i(0)}{N_{Control}} = \frac{525}{5} - \frac{442}{5} = 16.6$$

Here, $N_{Treatment}$ *and* $N_{Control}$ *are the numbers of employees in the respective groups. This estimation indicates an average productivity increase of 16.6 units due to the training, assuming no confounding factors.*

In terms of causality, the potential outcomes framework helps address issues like Simpson's paradox by structuring and quantifying causal effects. To navigate the paradox and avoid misleading conclusions, it's crucial to account for confounding variables and stratify analyses where necessary. This topic is discussed in depth in Section 2.9.2, especially in relation to Pearl's graphical methods.

Comparison with Pearl's Graphical approach. As we introduced in earlier sections, Pearl's graphical approach utilises DAGs to depict causal relationships, and it introduces the do-calculus for reasoning about interventions and counterfactuals. Here, we draw parallels between these two frameworks to give an idea of the approaches and advantages of each.

- **Representation**: While potential outcomes focus on the outcomes of treatments without specifying the underlying causal structure, Pearl's DAGs provide a visual representation of the causal mechanisms.
- **Intervention**: Both frameworks handle interventions, but Pearl's approach offers the do-calculus to compute interventional distributions, providing more explicit tools for reasoning about interventions.
- **Counterfactuals**: Pearl's formulation gives a clear mechanism to reason about counterfactuals through structural equations, whereas the potential outcomes framework considers counterfactuals implicitly.
- **Assumptions**: Both require untestable assumptions, such as the Stable Unit Treatment Value Assumption (SUTVA) in potential outcomes, and causal sufficiency in Pearl's framework.

A.2.3 Information Theoretic Approach

The information theoretic approach to causal inference, as detailed in Peters et al. [196], offers a distinctive perspective compared to Pearl's graphical framework and the Rubin-Causal Model. Unlike these models, which primarily utilise structural equations and potential outcomes, the information theoretic approach hinges on the statistical properties of observed variables, integrating principles of information theory into the study of causality.

This approach, fundamentally different in its reliance on statistical dependence measures, posits that causal relationships are asymmetric and can be captured through these dependencies. This asymmetry is formalised through concepts like the *independence of mechanisms* or the *algorithmic independence of conditionals* [204], suggesting that causal mechanisms operate independently from the distributions of the variables involved.

Key elements of the framework include:

- **Entropy:** The uncertainty of a random variable X is quantified by entropy:

  $$H(X) = -\sum_{x} p(x) \log p(x) = \mathbb{E}\left[-\log p(X)\right] \tag{A.3}$$

  This measure is foundational in identifying how much information is gained when one variable is known, impacting the inference of causality.

- **Mutual Information:** This quantifies the dependency between variables X and Y:
$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{A.4}$$
It's instrumental in discerning correlations that might hint at causal connections.

- **Causal Direction:** Employing entropy and mutual information, the causal direction can be inferred, often using additive noise models [199]:
$$Y = f(X) + N \tag{A.5}$$
Here, f represents a deterministic function, and N is noise independent of X.

The independence of mechanisms principle is crucial. It implies that the mechanism connecting cause and effect (e.g., f in $Y = f(X) + N$) is independent of the input distribution (the distribution of X). This principle helps distinguish causal relationships from mere statistical correlations, especially in complex, multivariate environments.

The framework's application extends from simple bivariate causal inference to multivariate settings. For a comprehensive introduction, see Peters et al. [23] and Janzing and Schölkopf [203].

In the context of ML and RL, this approach could offer novel pathways for embedding causality into learning algorithms. While not discussed extensively in this thesis, the information theoretic approach's potential for enhancing predictive models and decision-making processes in these fields warrants attention.

Example 17 (Inferring Causal Direction with Additive Noise Model)**.** *Consider a causal system with variables X and Y. Under the additive noise model (ANM), if X causes Y, then Y can be expressed as a function of X with additive noise N_y:*

$$Y = f(X) + N_y \tag{A.6}$$

Assume X is uniformly distributed over $[-1, 1]$ and N_y is standard normal noise. We examine both forward ($X \to Y$) and reverse ($Y \to X$) causal directions:

1. ***Model the Forward Direction** ($X \to Y$): Fit the model $Y = aX + N_y$ and estimate a. Test for independence between residuals and X using criteria like the Hilbert-Schmidt Independence Criterion. Independence supports $X \to Y$.*

2. ***Model the Reverse Direction** ($Y \to X$): Fit $X = bY + N_x$ and estimate b. Test for independence between residuals and Y. Lack of independence challenges $Y \to X$.*

If residuals are independent in the forward but not the reverse direction, it suggests $X \to Y$. However, independence in both directions or dependence in both directions complicates causal inference, underscoring the need to consider model assumptions and potential hidden confounders.

A.2.4 Causal Quantities

Understanding the specific causal quantities is crucial for any study involving causal inference. The Average Treatment Effect (ATE) is often the starting point, providing a population-level measure of the expected outcome difference between treated and untreated groups. Formulated as ATE $= \mathbb{E}[Y(1) - Y(0)]$, it lays the groundwork for other, more granular causal metrics.

Building upon the ATE, we turn our attention to Individual Treatment Effects (ITE), which focus on the treatment impact on specific units. This is denoted as $\tau_i = Y_i(1) - Y_i(0)$. ITE becomes particularly important when personalisation is key, such as in personalised medicine or recommendation systems.

Another refinement of ATE is the Average Treatment Effect on the Treated (ATT), represented as ATT $= \mathbb{E}[Y(1) - Y(0)|T = 1]$. ATT is invaluable when the focus is only on the subset of the population that actually receives the treatment, thereby narrowing the scope of investigation.

From the specificity of ATT and ITE, we extend to the Total Effect (TE), which captures not just the direct but also the cascading impacts of an intervention. This measure is pertinent in sociological or economic domains where indirect causal pathways are often significant.

Finally, we consider the Conditional Average Treatment Effect (CATE), represented by CATE$(X) = \mathbb{E}[Y(1) - Y(0)|X]$. CATE is crucial when we aim to condition the treatment effects on observed covariates, thereby investigating how the effect varies based on other variables.

Having established these causal quantities, the next logical question is how to actually identify or estimate them. This leads us to the issue of identifiability.

Table A.1: Mathematical Formulations and Levels of Application of Important Causal Quantities

Causal Quantity	Mathematical Formulation	Level of Application
ATE	$\mathbb{E}[Y(1) - Y(0)]$	Population-level
ITE	$\tau_i = Y_i(1) - Y_i(0)$	Individual-level
ATT	$\mathbb{E}[Y(1) - Y(0)|T = 1]$	Treated population
TE	Varies; includes intermediates	Population-level
CATE	$\mathbb{E}[Y(1) - Y(0)|X]$	Conditional on covariates

Table A.2: Positive and Negative Aspects of Important Causal Quantities

Causal Quantity	Positive Aspects	Negative Aspects
ATE	Simple; Broadly applicable	Masks individual heterogeneity
ITE	Personalised; Targeted interventions	Requires rich data; Computationally intensive
ATT	Focus on treated individuals; Practical	Biased if uncontrolled

Causal Quantity	Positive Aspects	Negative Aspects
TE	Captures cascading effects; Comprehensive	Complex; Requires full causal pathways
CATE	Addresses heterogeneity; Tailored interventions	Data-intensive; Model-dependent

A.3 Adjustment Criteria

The concept of the Back-Door Criterion is important in causal inference for establishing a causal relationship between variables. This concept, initially proposed by Pearl [193], helps in identifying the causal effect $P(y \mid \hat{x})$ for a given set of variables $Z \subseteq V$.

Definition A.3.1 (Back-Door Criterion)**.** *A variable set Z adheres to the back-door criterion for a pair of variables (X, Y) within a Directed Acyclic Graph (DAG) G, if the following conditions are met:*

(i) Z contains no nodes that are descendants of X.

(ii) All paths from X to Y, which have an arrow entering X, are obstructed by Z.

This criterion essentially implies that only the paths originating from X and leading to Y are allowed.

Theorem A.3.1 (Back-Door Adjustment)**.** *The identifiable causal effect of Z on Y, provided Z meets the back-door criterion with respect to (X, Y), is quantifiable. The causal effect can be calculated using the formula: $P(y \mid \hat{x}) = \sum_z P(y \mid x, z)P(z)$.*

The primary objective is to identify a suitable set of variables that, when conditioned upon, obstruct all spurious associations between a given variable or variables X and Y. One practical approach is to *intervene* on a subset of these variables, effectively eliminating the incoming edges to the independent variable X. For instance, an intervention such as $do(X = x)$ results in a distribution where the causal pathway $X \to M \to Y$, mediated by M, remains the sole connecting path between X and Y. The goal is to achieve this path blocking using observational data, avoiding the need for direct intervention which might be restricted due to ethical or financial considerations.

Example 18 (Back-door Adjustment)**.** *Imagine a causal model represented by a DAG with four variables X, Y, Z_1, and Z_2. In this model, Z_1 is a confounder influencing both X and Y, while Z_2 is a mediator impacted by X and influencing Y.*

In this setup, Z_1 represents the minimal set blocking all back-door paths from X to Y as per the back-door criterion. Since Z_2 is a descendant of X, it is excluded from the adjustment set. Adjusting for Z_1 enables the identification of the causal effect of X on Y, calculated as:

$$P(y \mid \hat{x}) = \sum_{z_1} P(y \mid x, z_1)P(z_1)$$

Here, Z_1 fulfills the back-door criterion, facilitating the estimation of the causal impact of X on Y.

Aside from the back-door criterion, the front-door criterion serves as another method for causal effect identification. It utilises the structure of intermediary variables. If a causal effect of X on Y operates through a mediator M, analysing the effect of X on M and subsequently M on Y should reveal the causal relationship between X and Y, assuming that any non-causal or back-door influences are appropriately controlled.

Definition A.3.2 (Front-Door Criterion)**.** *The front-door criterion is met by a variable set Z for the ordered pair (X, Y) when:*

(i) All direct paths from X to Y are intercepted by Z.

(ii) X to Y has no unobstructed back-door paths.

(iii) X blocks all back-door paths from Z to Y.

Theorem A.3.2 (Front-Door Adjustment)**.** *Given that Z satisfies the front-door criterion for (X, Y) and $P(x, z) > 0$, the causal effect of X on Y is identifiable. It can be computed using: $P(y \mid \hat{x}) = \sum_z P(z \mid x) \sum_{x'} P(y \mid x', z)P(x')$.*

A.4 Causal Structure Learning Algorithms

Inferred Causation (IC) Algorithm Originally introduced by Verma and Pearl [29], the IC algorithm operates on the premise of utilising v-structures to differentiate between different Markov equivalence classes of directed acyclic graphs (DAG). The algorithm predominantly identifies such v-structures to partially orient the edges of the graph. Like the PC algorithm (introduced below), the IC algorithm also relies on the Causal Markov and Faithfulness conditions. However, its distinction lies in its approach to the orientation phase, which puts emphasis on the identification and orientation of v-structures.

Peter-Clark (PC) Algorithm PC [119] is an algorithm that is very commonly used in practise because it is relatively fast and is readily available in many causal inference packages (e.g. [191, 192]). PC starts by assuming a complete graph, and iteratively reduces the number of edges until convergence to a Markov Equivalence Class. Given the correct assumptions and a large enough sample size, the PC algorithm is guaranteed to converge. The main assumptions are the Causal Markov and Faithfulness assumptions, i.i.d. samples, and no unmeasured confounders.

Generative Model Approach Here we provide pseudocode for a possible approach one could take for applying generative models to learn a causal model. This accompanies the discussion in the main text (Section 2.11).

Once the GAN is adequately trained, it can be used as a proxy for the data-generating process. We can then employ interventions on this model to simulate potential causal relationships.

Linear non-Gaussian Acyclic Model (LiNGAM) LiNGAM is a causal discovery method designed specifically for situations where the data are linear and acyclic but not necessarily Gaussian [200]. In typical structural equation models (SEM), the presence of non-Gaussian distributions can hinder the precise identification of the causal structure. However, LiNGAM, by leveraging the non-Gaussian nature of the data, enables

Algorithm 12: IC algorithm. Adapted from [29].

```
Input: Dataset $D$ with variable set $\mathbf{V}$, and a chosen significance level $\alpha$
Output: Partially directed graph $G$ with edge set $\mathbf{E}$
Initialise a complete undirected graph using variables in $\mathbf{V}$
for each non-adjacent variable pair $(X, Y)$ in $G$ do
    for each subset $\mathbf{S}$ of $\mathbf{V} \setminus \{X, Y\}$ do
        if Conditional independence test $I(X, Y|\mathbf{S})$ is not significant at level $\alpha$ then
            Remove the edge between $X$ and $Y$ from $G$
            Record $\mathbf{S}$ as the separating set for $(X, Y)$
            Break the inner loop and continue with the next pair of non-adjacent variables
Apply orientation rules to identify v-structures:
for each triplet $(X, Y, Z)$ in $G$ where $X - Y$ and $Y - Z$ but not $X - Z$ do
    if $Y$ is not in the separating set for the pair $(X, Z)$ then
        Orient $X - Y - Z$ as $X \to Y \leftarrow Z$
Apply additional orientation rules based on the directed edges and identified v-structures.
```

Algorithm 13: PC algorithm. Adapted from [206].

```
Input: Dataset $D$ encompassing a variable set $\mathbf{V}$, and a chosen significance level $\alpha$
Output: Partially directed acyclic graph $G$ featuring edge set $\mathbf{E}$
Begin with a complete undirected graph using nodes from $\mathbf{V}$
Set depth $d = 0$
repeat
    for each pair of adjacent nodes $X$ and $Y$ in $G$ do
        if $|adj(X, G) \setminus \{Y\}| \geq d$ then
            FoundIndependence ← false
            for each subset $Z \subseteq adj(X, G) \setminus \{Y\}$ with $|Z| = d$ do
                Conduct test $I(X, Y|Z)$
                if Independence $I(X, Y|Z)$ is found at level $\alpha$ then
                    Remove edge between $X$ and $Y$ from $G$
                    Record $Z$ as the separator for $(X, Y)$
                    FoundIndependence ← true
                    break
            if !FoundIndependence then
                Continue with the next adjacent node pair
    Increment depth $d$ by 1
until no changes are made to $G$;
Execute edge orientation rules to infer directed edges (consider using Meek's rules).
```

Algorithm 14: Causal discovery using GANs.

```
Input: Dataset $D$ with variables $\mathbf{V}$, GAN architecture with generator $G$ and
       discriminator $D$, number of training epochs $T$, learning rate $\eta$
Output: The causal structure inferred from interventions on the generator
Initialise the weights of the generator $G$ and the discriminator $D$ randomly
for $epoch = 1$ to $T$ do
    for each batch in $D$ do
        Generate fake data $\mathbf{V}_{fake} = G(\mathbf{z})$ using random noise $\mathbf{z}$
        Calculate discriminator loss $L_D$ (e.g., binary cross-entropy) for real and
          fake data
        Update weights of $D$ to minimise $L_D$ using learning rate $\eta$
        Generate fake data $\mathbf{V}_{fake} = G(\mathbf{z})$ using random noise $\mathbf{z}$
        Calculate generator loss $L_G$ (e.g., binary cross-entropy) using
          discriminator's feedback
        Update weights of $G$ to minimise $L_G$ using learning rate $\eta$
for each variable $v_i$ in $\mathbf{V}$ do
    Perform an intervention, such as setting $v_i$ to a specific value $v'$
    Generate intervened data $\mathbf{V}_{intervened} = G_{do(v_i=v')}(\mathbf{z})$
    for each other variable $v_j \neq v_i$ in $\mathbf{V}$ do
        Calculate the effect of the intervention on $v_j$ (e.g., using statistical tests
          or effect size measures)
        if the change in the distribution of $v_j$ is significant according to a
          predefined threshold then
            Mark $v_i$ as a potential cause of $v_j$
```

a unique determination of the causal order. The algorithm operates under the premise that each observed variable is a linear function of its direct causes and a non-Gaussian noise term. The central idea is to exploit the unique statistical characteristics of non-Gaussian distributions. Specifically, while mixtures of non-Gaussian variables gravitate towards Gaussianity, their demixtures do the opposite. This property, known as the non-Gaussian nature of the data, aids in revealing the causal ordering of variables.

A.5 Interplay Between Structural Causal Models and Differential Equations

Although tangential to the thesis's main focus, this subsection provides crucial insights by bridging SCMs and differential equations—two frameworks foundational to scientific theory. Despite a rising academic interest in causal models, a conceptual chasm remains between traditional differential equations and contemporary causal frameworks.

We direct our discussion to the specific alignment between SCMs and Ordinary Differential Equations (ODE) [222]. Although the scope of this alignment has been expanding in the literature to include other types of differential equations, we limit our discussion to ODEs. Notably, the Pearlian SCM framework enforces acyclic constraints on its graphical models. While beneficial for isolating causal pathways among variables,

Algorithm 15: LiNGAM algorithm. Adapted from [200].

```
Input: Dataset $D$ with a set of variables $\mathbf{V}$
Output: Causal ordering of the variables in $\mathbf{V}$
Estimate the mixing matrix $\mathbf{B}$ from data $D$ using independent component
  analysis (ICA) Compute the connection strengths for each pair of variables
  $(X, Y)$ based on the entries in $\mathbf{B}$
Initialise an empty directed graph $G$ with nodes $\mathbf{V}$ and no edges
for each pair of variables $(X, Y)$ do
    if connection strength$(X, Y)$ exceeds the predefined significance threshold
      then
        Add directed edge $X \to Y$ to $G$ based on the sign of the connection
          strength
Derive the causal order from the directed acyclic graph $G$ by performing a
  topological sort of the nodes
```

these constraints fall short in capturing systems with feedback mechanisms, like coupled harmonic oscillators.

Modelling Cyclicity Mooij et al. [222] address this limitation. They assert that SCMs can seamlessly adapt to feedback systems by relaxing the acyclicity constraints. Through the lens of underlying ODE systems, SCMs can be reinterpreted to illuminate how interventions on variables affect the system's equilibrium state.

We consider a dynamical system $\mathcal{D}$ consisting of D coupled first-order ODEs, initiated with $X_0 \in \mathcal{R}_{\mathcal{I}}$. The set of variable labels is defined as $\mathcal{I} = \{1, \ldots, D\}$. Mathematically, $\dot{X}_i(t) = f_i(\boldsymbol{X}_{pa_{\mathcal{D}}(i)}), X_i(0) = (\boldsymbol{X}_0)_i \ \forall i \in \mathcal{I}$. Here, $\dot{X}_i(t)$ represents the time derivative of X_i, $pa_{\mathcal{D}}(i)$ indicates the parent variables, and f_i maps the parent variables to X_i. We can graphically represent this system's *structure*, similar to SCMs. The Lotka-Volterra model serves as a canonical example, especially pertinent when exploring multi-agent interactions, as further elaborated in section 3.

Lotka-Volterra model example Let's consider the Lotka-Volterra model, introduced in 1910 by Alfred J. Lotka [276] and later commonly used for modelling predator-prey dynamics [277]. This seemingly simple model produces remarkably interesting interactions. Let's assume we are working with the predator-prey analogy where we start with an abundance of prey, X_1 and predators X_2. We are particularly interested in how the populations change over time due to the interactions of the populations as well as time. The variables at play are as follows:

1. X_1 represents the number of prey.
2. X_2 represents the number of predators.
3. X_1 and X_2 represent the growth rates of the respective populations with respect to time t.
4. $\theta_{ij} \quad i, j \in \{1, 2\}$ are model parameters controlling interaction of the populations.

$$\begin{cases} \dot{X}_1 = X_1(\theta_{11} - \theta_{12}X_2) \\ \dot{X}_2 = -X_2(\theta_{22} - \theta_{21}X_1) \end{cases} \qquad \begin{cases} X_1(0) = a \\ X_2(0) = b \end{cases}$$

This dynamical system can be implemented simply in Julia using the Differential Equations package. The code is as follows:

```
using DifferentialEquations

# Define the DE model
function lotka_volterra(du,u,p,t)
    x, y = u
    a, b, delta, gamma = p
    du[1] = dx = a*x - b*x*y
    du[2] = dy = -delta*y + gamma*x*y
end

# Define initial conditions and timespan
tspan = (0.0,10.0)
p = [1.5,1.0,3.0,1.0]
prob = ODEProblem(lotka_volterra,[u1,u2],tspan,p)

# Solve and plot the DEs
sol = solve(prob)
using Plots
plot(sol)
```

A.5.1 Graph Distance Evaluation

Comparing the performance of various causal discovery algorithms necessitates dependable metrics to gauge the precision and quality of the produced causal graphs. Given that these algorithms yield graphical representations of causality, their evaluation becomes indispensable. Among the tools for such assessments are graph distance metrics, with the choice of metric often relying on the problem specifics and the available ground truth.

One of the commonly adopted metrics is the **Area Under the Precision-Recall Curve** (AUPRC/AUC-PR). It offers a comprehensive view of performance by illustrating how the algorithm fares across different thresholds. By juxtaposing precision, which refers to the accuracy of the predicted edges, against recall, which denotes the identification rate of true edges, a clearer picture emerges. A higher AUPRC means the algorithm strikes a better balance between precision and recall. Notably, when classes display imbalance, the AUPRC stands out by accounting for both false positives and negatives.

On a more structural note, the **Structural Hamming Distance** (SHD) quantifies the variance between two graphs. Essentially, it enumerates the operations (additions, deletions, or reversals of directed edges) required to align the predicted graph with the true one. A lower SHD signals a closer match to the true causal graph, and if the SHD amounts to zero, it means the two graphs are indistinguishable.

When the emphasis is on interventions, the **Structural Intervention Distance** (SID) [285] becomes vital. It gauges the disparity between predicted interventions on the inferred graph and the true interventions on the original. A diminished SID indicates that predicted interventions resonate more with the genuine causal relationships. This metric stands out when the intent of causal discovery leans toward directing interventions, as it contrasts predicted outcomes against actual results.

Another important metric is the **False Discovery Rate** (FDR), which captures the ratio of incorrect edges to the total predicted edges. A decline in FDR translates to fewer mistakes in the discoveries. If an algorithm exhibits a high FDR, it might be prone to overfitting or pinpointing random relationships, underscoring the importance of maintaining a low FDR to uphold the trustworthiness of causal findings.

Lastly, the **F1 Score** emerges as the harmonic mean of precision and recall. This metric, especially pertinent when there's an imbalance in classes, guarantees an even-handed evaluation. A soaring F1 score implies the algorithm adeptly balances precision and recall, and when the score maxes out at 1, the algorithm is faultless. Such a score becomes invaluable when neither precision nor recall should overshadow the other, making it ideal when high true positive rates matter.

In practice, these (among other) metrics might see individual or combined usage, contingent on the study's aims. The paramount task lies in aligning chosen metrics with research goals. For instance, if the priority is predicting accurate interventions, SID becomes more consequential. Conversely, if the focus sways towards the graph's aggregate accuracy, then SHD or the F1 score might take precedence. Conventionally, a blend of these metrics ensures a thorough assessment of the algorithm's competence.

A.6 Causality-like Methods

While the causal frameworks primarily championed by pioneers like Judea Pearl lay the foundational understanding of causal relationships, many methods in ML and statistics offer insights that, while not strictly causal in nature, can resemble or hint at causal dynamics. These methods often exploit strong correlations or structured relationships among variables to deduce patterns that might be interpreted in a causal-like manner. It is imperative to approach the results from these methods with caution, recognising their limits and understanding that correlation does not necessarily imply causation [294].

Correlation-based Approaches are both common and noteworthy. In general, these methods aim to quantify the strength and direction of linear relationships between two variables [293]. Many of these statistical methods make use of Pearson's correlation coefficient. Though statisticians would not consider this to be causal, the causal misinterpretation of correlation coefficients are common. While a high correlation might suggest a possible causal link, it's vital to remember it doesn't account for confounding factors or guarantee a cause-effect relationship.

Another set of techniques involve **Decision Trees and Random Forests**. As ML models, they can implicitly capture and visualise relationships between variables [292, 295]. For instance, a variable that frequently appears at the top of decision trees might be viewed as 'important' or even a root cause for certain outcomes. But, again, this isn't indicative of a strictly causal relationship.

The domain also benefits from **Feature Importance from Machine Learning Models**. Techniques such as SHAP (SHapley Additive exPlanations) or permutation feature importance play a role in ranking features based on their impact on a model's prediction [296]. Although a high importance suggests that a variable strongly influences the outcome, a clear causal interpretation is lacking.

From the field of econometrics, **Granger Causality** is often employed [299, 300]. It tests whether past values of one time series can predict another. While it identifies a type of 'predictive causality' based on temporal precedence, it doesn't establish causation in the Pearlian sense.

Information Theoretic Approaches like Transfer Entropy also deserve mention [297, 301]. Such methods, especially relevant for time series, hint at causal dynamics by gauging how much uncertainty in predicting one variable's future states is reduced by understanding another variable's past states.

Lastly, **Regression-based Approaches**, such as linear regression, are frequently used to quantify relationships between predictors and outcomes [298, 302]. However, extracting causal implications from them requires more than just observing coefficient estimates. One must control for confounders and ensure other criteria are met to move beyond mere associations.

In sum, while causality-like methods illuminate relationships between variables, it's crucial to interpret their results carefully. Using these methods with strictly causal methods might be one way to present a balanced view of the data. That is, using both associative patterns and genuine causal relationships.

A.7 Motivating Examples for Application of Causal Inference

In relation to Chapter 2, this section further clarifies the distinction between statistical correlations and causality through emblematic examples. These scenarios highlight situations where conventional statistical reasoning might mislead, while causal interpretations offer clarity.

Example 19 (Simpson's Paradox)**.** *Simpson's Paradox is a well-known example of complexities in statistical analysis, where trends in different groups change or disappear when the groups are combined.*

Consider a hypothetical example of a university's graduate admissions, appearing to show gender bias. At the departmental level:

- ***Department A:*** *90% of male applicants gain admission versus 80% of their female counterparts.*
- ***Department B:*** *A mere 10% of male applicants are admitted as compared to 20% of female applicants.*

Looking at these numbers separately, Department A seems to favour males, while Department B appears to favour females. However, if we combine the data and assume *more females apply to Department A (which has a high acceptance rate) and more males apply to Department B (which is more selective), the overall data might show that females have a higher acceptance rate.*

Simple statistical analysis might suggest the university is biased based on overall data. However, causal inference reveals a deeper issue: the choice of department is a confounding variable that affects the apparent trend.

Key Takeaway: The example illustrates how causal inference can reveal underlying factors in observed trends that are not apparent in simple statistical analyses.

Example 20 (Berkson's Paradox)**.** *Berkson's Paradox, also known as Berkson's Bias, is a counterintuitive phenomenon in statistics and probability that challenges our intuition about correlation and causation.*

Imagine a study assessing the relationship between fitness levels and the frequency of hospital visits. In this study, two groups of individuals are considered: those who frequently visit the gym and those who are often admitted to the hospital.

- ***Gym-goers:*** *Individuals who regularly visit the gym tend to have fewer hospital admissions.*
- ***Non-gym-goers:*** *Individuals who rarely or never visit the gym have a mixed record of hospital admissions, with some having frequent visits and others very few.*

At first glance, the data might suggest a negative correlation between gym attendance and hospital visits, implying that those who frequently visit the gym are less likely to require hospital care. However, this interpretation overlooks a critical selection bias: the study does not account for healthy individuals who neither visit the gym frequently nor require hospitalisation often.

Berkson's Paradox reveals itself here: the absence of this 'healthy' group in the data skews the apparent relationship between gym attendance and hospital visits. Causal inference, considering this missing group, helps to correct the bias and leads to a more nuanced understanding that regular gym attendance does not necessarily cause fewer hospital visits.

Key Takeaway: This example highlights the importance of considering potential selection biases in data when drawing causal inferences, as overlooking such biases can lead to misleading conclusions.

Example 21 (Ice Cream Sales and Drowning Incidents)**.** *A dataset depicts a robust correlation between the sales of ice cream and incidents of drowning. A simplistic statistical inference might suggest that a surge in ice cream sales augments the risk of drownings—a patently implausible conclusion.*

Causal deduction, however, unravels the underlying truth: both these phenomena are modulated by an extraneous factor, namely temperature. As the mercury rises during sweltering summer months, there's a surge in ice cream sales. Simultaneously, the allure of swimming to beat the heat results in an amplified risk of drowning. The *temperature acts as a confounding variable, underscoring the nuances that causality can unearth which pure correlations might overlook.*

Example 22 (Thought Experiment in Reinforcement Learning)**.** *Though we develop and consider RL in much more detail in the next chapter, we consider a simple thought experiment here.*

Visualise a RL agent tasked with mastering a video game where the protagonist accrues coins whilst evading adversaries. The agent's modus operandi is governed by a reward function: accumulating coins amplifies its score, whereas confrontations with foes deplete it.

Over iterative training cycles, the agent discerns a pattern: a specific auditory warning invariably precedes a negative reward. A myopic statistical model might deduce that the auditory cue is detrimental, leading the agent to eschew actions triggering this sound—even when they might be strategically advantageous.

Contrastingly, a causal paradigm enables the agent to comprehend that the auditory signal is merely an indicator, not the causative factor of the negative reward. The real menace is the lurking enemy. Recognising this cause-effect relationship—the auditory cue heralds an impending adversary which can be strategically outmaneuvered—the agent can optimise its strategy, leveraging the cue as a beneficial sentinel rather than an intrinsic threat. This vignette accentuates the centrality of causality in decision-making paradigms, transcending even artificial milieus.

Appendix B

RL Appendix

This appendix complements the main discussion on RL in Chapter 3 with a focus on advanced methodologies and algorithms. The structure of this chapter follows the order in which references are made to this appendix from the main body of this thesis. The aim is to offer additional details, proofs, or algorithms that might not directly contribute to the main argument of the thesis.

B.1 Multi-Armed Bandits

Multi-armed bandits (MAB) are a simplified framework in RL, focusing on the exploration-exploitation trade-off in decision-making. In MAB problems, a learner interacts with an environment consisting of several actions (or arms), each providing a stochastic reward. The learner's goal is to maximise the cumulative reward over time.

The mathematical formulation of MAB can be focused on the concept of regret, which is defined as the difference between the reward obtained by the optimal strategy and the strategy followed by the learner. Formally, if $R_t(a)$ is the reward received at time t from arm a, and a^* is the optimal arm, the regret after T trials is:

$$\text{Regret}(T) = \sum_{t=1}^{T} R_t(a^*) - R_t(a_t)$$

where a_t is the arm chosen at time t.

The exploration-exploitation dilemma, fundamental in MAB, is also a core aspect of general RL. MAB's insights are vital for understanding and developing strategies in more complex RL environments.

B.2 Optimisation

While the basic gradient descent approach is foundational, its stochastic variant, Stochastic Gradient Descent (SGD), and advanced optimisers like Adam [185], have become the mainstay in training deep NNs due to their efficiency in computing gradients, essential for parameter updates.

Beyond gradient-based methods, the broader optimisation landscape encompasses tech-

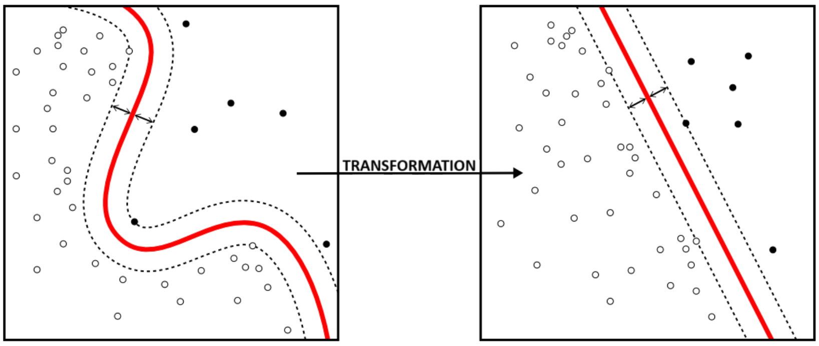


Figure B.1: Transformation of a dataset with a non-linear boundary into one that is linearly separable.

niques such as conjugate gradient descent and the Broyden-Fletcher-Goldfarh-Shanno algorithm (BFGS) [186]. Although less dominant in the deep learning domain, these methods hold significance in specific applications. The rise of NNs as a primary model in deep learning necessitates a deep understanding of parameterisation. Given the vastness of parameters in these networks and their interactions, efficient optimisation methods are crucial for effectively training them. Parameterisation, gradients, and the ensuing quest for optimality collectively contribute to the success of NNs in myriad applications. Listing B.1 is a concise Python implementation illustrating gradient descent. Following standard gradient descent implementations, we have some function f and an associated gradient function df which is the gradient in the usual sense, $\frac{df}{dx}$ (assuming univariate). $x0$ is the initial value for x, and lr is the learning rate chosen for this implementation. We can select the number of iterations, n_{iter}, to run the algorithm until it is 'close' to convergence.

Listing B.1: Implementation of the gradient descent algorithm in Python.

```python
def gradient_descent(f, df, x0, lr=0.01, n_iter=1000):
    '''Numpy implementation of gradient descent.'''
    x = x0
    for i in range(n_iter):
        x = x - lr * df(x)
    return x
```

B.3 Deep Q-learning

Deep Q-learning (DQN) employs a Huber loss function to stabilise learning by treating the loss quadratically for small deviations and linearly for large ones, preventing substantial fluctuations in Q-value approximations. A comprehensive pseudocode of DQN can be found in Algorithm 16.

B.4 Direct Policy Differentiation

This section develops the mathematical foundations of direct policy optimisation in RL, a method crucial for efficiently learning policies in various contexts. This approach is

Algorithm 16: Deep Q-Learning algorithm.

```
Input: Replay memory capacity $N$, number of episodes $M$, number of steps
       $T$, discount factor $\gamma$, exploration probability $\epsilon$, update frequency $C$
Output: optimised policy $\pi^*$ based on the learnt $Q$
Initialise replay memory $D$ to capacity $N$
Initialise action-value function $Q$ with random weights $\theta$
Initialise target action-value function $\hat{Q}$ with weights $\theta^- = \theta$
for episode = 1 to M do
    Initialise sequence $s_1 = \{x_1\}$ and pre-processed sequence $\phi_1 = \phi(s_1)$
    for $t$ = 1 to $T$ do
        if $Random() < \epsilon$ then
            Select random action $a_t$
        else
            Select $a_t = \text{argmax}_a Q(\phi(s_t), a; \theta)$
        Observe reward $r_t$ and image $x_{t+1}$
        Set $s_{t+1} = s_t, a_t, x_{t+1}$ and pre-process $\phi_{t+1} = \phi(s_{t+1})$
        Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $D$
        Sample a random sub-sample of transitions from $D$
        Compute $y_j$: if episode terminates at step $j + 1$ then
            $y_j = r_j$
        else
            $y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$
        Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to $\theta$
        Reset $\hat{Q} = Q$ every $C$ steps
Return the optimal policy $\pi^*$ based on the learnt $Q$.
```

powerful due to its ability to handle high-dimensional action spaces and continuous domains.

B.4.1 Mathematical Derivation

The key to understanding direct policy differentiation lies in the manipulation of the gradient of the expected reward function, $J(\theta)$, with respect to the policy parameters, θ. The derivation starts by expressing the gradient as an integral over all possible trajectories τ:

$$\begin{aligned}
\nabla J(\theta) &= \int \nabla_\theta \pi_\theta(\tau) r(\tau) d\tau \\
&= \int \pi_\theta(\tau) \frac{\nabla_\theta \pi_\theta(\tau)}{\pi_\theta(\tau)} r(\tau) d\tau \\
&= \int \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) r(\tau) d\tau \\
&= E_{\tau \sim \pi_\theta(\tau)} \left[\nabla_\theta \log \pi_\theta(\tau) r(\tau) \right]
\end{aligned}$$

Here, $\pi_\theta(\tau)$ denotes the probability of trajectory τ under the policy parameterised by

θ, and $r(\tau)$ represents the reward associated with τ.

B.4.2 Finite-Horizon Scenario

In a finite-horizon setting, where the trajectory has a defined end, the policy for a trajectory τ comprising states s_t and actions a_t over a time horizon T is given by:

$$\begin{aligned}
\pi_\theta(\tau) &= p(s_1)\prod_{t=1}^{T}\pi_\theta(a_t|s_t)p(s_{t+1}|s_t,a_t)\\
\log\pi_\theta(\tau) &= \log p(s_1) + \sum_{t=1}^{T}\log\pi_\theta(a_t|s_t) + \log p(s_{t+1}|s_t,a_t)\\
\nabla_\theta \log\pi_\theta(\tau) &= \sum_{t=1}^{T}\nabla_\theta\log\pi_\theta(a_t|s_t)
\end{aligned}$$

B.4.3 Practical Policy Update

The gradient of the objective function, $\nabla J(\theta)$, is estimated through sampling, leading to the following practical policy update rule:

$$\begin{aligned}
\nabla J(\theta) &= \mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\left(\sum_{t=1}^{T}\nabla_\theta\log\pi_\theta(a_t|s_t)\right)\left(\sum_{t=1}^{T}r(s_t,a_t)\right)\right]\\
&\approx \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta\log\pi_\theta(a_t|s_t)\right)\left(\sum_{t=1}^{T}r(s_t,a_t)\right)
\end{aligned}$$

This leads to the update rule for the policy parameters:

$$\theta \longleftarrow \theta + \alpha\nabla J(\theta),$$

where α is the learning coefficient. This update rule is applicable even in partially observed MDPs, showcasing the flexibility of direct policy differentiation in various RL contexts.

B.5 MARL Algorithms

Multi-agent reinforcement learning (MARL) involves various algorithms designed to tackle the complexities of learning in environments with multiple agents. These algorithms address challenges such as non-stationarity, large action spaces, and credit assignment. Key MARL algorithms include:

Independent Q-Learning (IQL) [268]: In IQL, each agent independently applies Q-learning, treating other agents as part of the environment. While simple, IQL struggles with non-stationarity as it does not account for the learning and adaptation of other agents.

Joint Action Learning (JAL) [267]: JAL extends the learning process by considering the actions of all agents when updating Q-values, better handling non-stationarity but suffering from the curse of dimensionality as the number of agents increases.

Value Decomposition Networks (VDN) [266]: VDNs decompose the joint action-value function into individual value functions for each agent, allowing for cooperation and manageable learning complexity.

Counterfactual Multi-Agent (COMA) Policy Gradients [264]: COMA addresses the credit assignment problem in cooperative settings using a centralised critic that evaluates each agent's action contribution to the overall performance.

Mean Field Multi-Agent Reinforcement Learning [263]: This method simplifies interactions in large-scale agent scenarios by considering the average effect of neighbouring agents, enhancing scalability and reducing computational complexity.

Each algorithm has its strengths and is suited for different types of multi-agent environments and objectives.

Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [265]: MADDPG, an extension of DDPG, uses a centralised critic for policy evaluation and a decentralised actor for action execution, facilitating information sharing during training while preserving independent decision-making.

MADDPG supports both cooperative and competitive interactions among agents. MADDPG utilises centralised learning with decentralised execution, allowing agents to benefit from shared information during training while acting independently during execution.

B.6 Uncertainty Estimation in Model-Based RL

In the context of MBRL, ensuring the accuracy and reliability of predictive models is important. One potential avenue for achieving this reliability is through the estimation of uncertainty. This uncertainty serves as an indicator, highlighting areas where the model might be deficient, thereby reducing the need for extensive data collection. The capturing of uncertainty information could mean the difference between an agent taking a risky action with unintended consequences, versus one that is calculated and safe.

The concept of uncertainty in this context can be bifurcated into two distinct types:

1. Aleatoric (statistical) uncertainty: Inherent noise in the data itself.
2. Epistemic (model) uncertainty: Uncertainty related to the model's parameters, given the available data.

A simplified analogy encapsulates the difference: "The model may be certain about the data, but our confidence in the model can vary." The notion of high entropy being synonymous with high uncertainty is misleading. In RL, there's a nuanced relationship between entropy and uncertainty. For instance, consider a model that overfits: such a model might exhibit low entropy despite being unreliable.

One method to understand and estimate uncertainty is the Bayesian NN. In this approach, each weight connecting neurons within the network is treated as a distribution,

Algorithm 17: MADDPG Algorithm for Multi-Agent Reinforcement Learning.

```
Input: A multi-agent environment with states $\mathcal{S}$, actions $\mathcal{A}_i$ for each agent $i$,
       transition probabilities $T$, reward functions $R_i$ for each agent, discount
       factor $\gamma$, and a replay buffer $\mathcal{R}$.
Output: Optimal policy functions $\mu_\star$ for each agent.
Initialise critic networks $Q_{\theta_i}$ and actor networks $\mu_{\phi_i}$ for each agent $i$.
Initialise target networks $Q_{\theta'_i}$ and $\mu_{\phi'_i}$ with weights $\theta'_i \leftarrow \theta_i$ and $\phi'_i \leftarrow \phi_i$.
Initialise replay buffer $\mathcal{R}$.
for episode = 1 to M do
    Receive initial observation state $\mathbf{o}^1$.
    for $t$ = 1 to $T$ do
        for each agent $i$ do
            $a_i^t \leftarrow$ Select action using $\mu_{\phi_i}$ with exploration noise added.
        Execute joint actions $\mathbf{a}^t$ and observe reward $r^t$ and new observation
          $\mathbf{o}^{t+1}$.
        Store transition $(\mathbf{o}^t, \mathbf{a}^t, r^t, \mathbf{o}^{t+1})$ in $\mathcal{R}$.
        for each agent $i$ do
            Sample a batch of transitions from $\mathcal{R}$.
            $L(\theta_i) \leftarrow$ Calculate critic loss based on sampled transitions and
              update the critic network.
            Update the actor network using the sampled policy gradient.
            Update the target networks:
                $\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i,$
                $\phi'_i \leftarrow \tau\phi_i + (1-\tau)\phi'_i.$
```

rather than a fixed value (see Figure B.2). Typically, these weights are considered independent, leading to the assumption that the distribution over the weights is the product of the individual distributions, expressed as $p(\theta|D) = \prod_i p(\theta_i|D)$, where each marginal follows a normal distribution, $p(\theta_i) = N(\mu_i, \sigma_i)$. While this assumption simplifies computation, it does introduce potential limitations, making it a common yet controversial practice in ML.

While Bayesian NNs remain an active area of research, another technique gaining traction in the quest for uncertainty estimation is the use of bootstrap ensembles. Here, multiple networks are trained on independent datasets, all aiming to predict the same policy. These disparate policies, when viewed collectively, can help approximate the full distribution, described as $p(\theta|D) \approx \frac{1}{N}\sum_i \delta(\theta_i)$. Essentially, a mixture model is formed using these models, leading to the relationship:

$$\int p(s_{t+1}|s_t, a_t, \theta)p(\theta|D)d\theta \approx \frac{1}{N}\sum_i p(s_{t+1}|s_t, a_t, \theta_i).$$

These independent datasets can be synthetically created by sampling with replacement from the primary dataset, D.

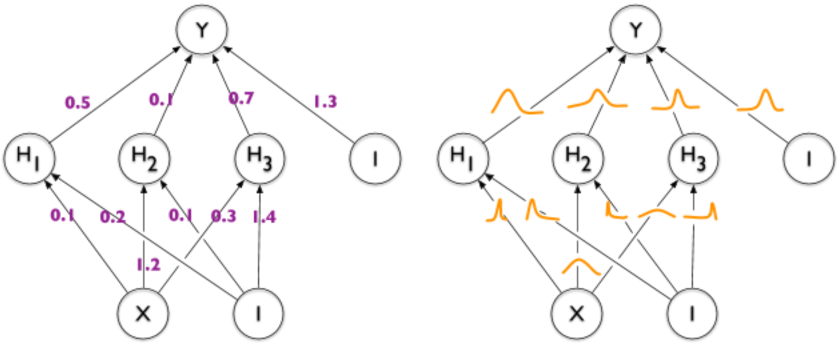


Figure B.2: Illustration of a Bayesian NN.

Appendix C

Causal RL Appendix

This appendix extends the discussion on causal RL presented in Chapter 4. The structure of this chapter follows the order in which references are made to this appendix from the main body of this thesis. The aim is to offer additional details, proofs, or algorithms that might not directly contribute to the main argument of the thesis.

C.1 Theoretical Foundations

This section presents key theoretical aspects of causal RL, starting with regret bounds in K-MAB problems.

Theorem C.1.1 (B-kl-UCB Regret Bounds)**.** *Consider a K-MAB problem with rewards bounded in* $[0, 1]$*, with each arm* $x \in \{1, \ldots, K\}$*, and expected reward* $\mu_x \in [l_x, h_x]$ *s.t.* $0 < l_x < h_x < 1$*. Taking* $f(t) = \log(t) + 3\log(\log(t))$*, in the B-kl-UCB algorithm (shown in algorithm 18), the number of draws of* $\mathbb{E}[N_x(T)]$ *for any sub-optimal arm a is upper bounded for any horizon* $T \geq 3$ *as:*

$$\begin{cases} 0 & \textit{if } h_x < l_{\max} \\ 4 + 4e\log(\log(T)) & \textit{if } h_x \in [l_{\max}, \mu^*) \\ \frac{\log(T)}{KL(\mu_x, \mu^*)} + \mathcal{O}\left(\frac{\log(\log(T))}{KL(\mu_x, \mu^*)}\right) & \textit{if } h_x \geq \mu^* \end{cases}$$

This is orchestrated by utilising prior knowledge to formulate a general SCM that aligns with all available models. A deeper examination of this concept is provided with a representation of a stochastic MAB problem with a prior depicted as a list of bounds over the expected rewards. For any given bandit arm x, the expected reward is bounded by $\mathbb{E}_{x \in \pi(u)}[Y \mid do(x)]$, with u being the contextual variable. The proposed B-kl-UCB algorithm (Algorithm 18), an augmentation of the well-known kl-UCB algorithm [79], leverages these bounds to expedite the learning process.

Algorithm 18: B-kl-UCB

```
Result: Compute causal bounds for non-identifiable transfer task
Input: Non-decreasing function $f : \mathbb{N} \rightarrow \mathbb{R}$
Input: A list of bounds over $\mu_x : \{[l_x, h_x]\}_{x \in \{1,\dots,K\}}$.
Exclude any arm $a$ with $h_x < l_{max}$. Let $K'$ denote the number of remaining
  arms. Pull each arm of $\{1, \dots, K\}$ once. for $t = K^{prime}$ to $T - 1$ do
    for each arm $x$ do
        Compute $\hat{U}_x(t) = \min\{U_x(t), h_x\}$ where
          $U_x(t) = \sup\{\mu \in [0, 1] : KL(\hat{\mu}_x(t), \mu) \leq \frac{f(t)}{N_x(t)}\}$.
    end
    Select an arm $X_t = \arg\max_{x \in \{1,\dots,K'\}} \hat{U}_x(t)$.
end
```

The promising connection between causal inference, transfer learning, and the generalisation of policy learning lays the foundation for a multitude of intriguing research directions. This integration could help RL agents to move beyond the limitations of their training environments, offering improved adaptability and learning efficiency.

C.2 Intervention Sets

Consider the information acquired by an agent engaged with an SCM-MAB as symbolised by $\langle G, Y \rangle$.

Definition C.2.1 (Minimal Intervention Set (MIS) [132]). *A subset of endogenous variables, denoted as $\boldsymbol{X} \subseteq \boldsymbol{V} \setminus \{Y\}$, is termed a minimal intervention set in relation to $\langle G, Y \rangle$ if no smaller subset, $\boldsymbol{X}' \subset \boldsymbol{X}$, exists that yields the same mean reward, i.e., $\mu_{x'} = \mu_x$ for every $x' \in \boldsymbol{X'}$, in every SCM adhering to the causal graph's stipulated structure.*

Reflecting on this definition, it becomes evident that intervening on the reward variable, Y's ancestors is both a necessary and sufficient condition for achieving an MIS.

Proposition C.2.1 (Minimality [132]). *A set $\boldsymbol{X} \subseteq \boldsymbol{V} \setminus \{Y\}$ qualifies as an MIS for causal graph G and reward variable Y if and only if $\boldsymbol{X} \subseteq an(Y)_{G_{\overline{X}}}$.*

This proposition furnishes a strategy for identifying MISs given the available information. A simple perusal of all possible subsets of endogenous variables $\boldsymbol{X} \setminus \{Y\}$ while checking the proposition suffices. However, the structure of causal graphs may dictate that certain variable sets are always superior for intervention, thereby establishing a partial ordering. This scenario propels the quest for formalising the concept of a possibly optimal MIS.

Definition C.2.2 (Possibly-Optimal MIS (POMIS) [132]). *Given the information $\langle G, Y \rangle$ and an MIS $\boldsymbol{X}$, a POMIS scenario arises if there exists a particular SCM adhering to causal graph G rules such that $\mu_{\boldsymbol{x}^*} > \mu_{\boldsymbol{z}^*} \forall \boldsymbol{Z} \in \mathbb{Z} \setminus \{\boldsymbol{X}\}$. Here, $\mathbb{Z}$ embodies the potential MISs compliant with G and Y.*

The preceding discussion of POMIS facilitates the recognition of optimal intervention points, highlighting the 'where' of allocating intervention.

C.3 Markovian Property in MDPUCs

Theorem C.3.1 (Markovian Property in MDPUCs [139]). *Given an MDPUC model $M\langle\gamma, U, X, Y, S, F, P(u)\rangle$, a policy $\pi \in F_{exp} = \{\pi \mid \pi : S \to X\}$ (state to action map), and an initial state $s^{(t)}$, the agent executes actions $do(X^{(t)} = x^{(t)})$ at round t and $do(X^{([t+1,t+k])} = \pi)$ subsequently ($k \in \mathbb{Z}^+$), the following relationship holds:*

$$P\left(Y^{t+k}_{x^{(t)},x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) = P\left(Y^{t+k}_{x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}\right)$$

Further extension of MDPUCs to encompass counterfactual policies is possible by considering $F_{ctf} = \{\pi \mid \pi : S \times X \to X\}$, a set of functions defining the relationship between the current state $s^{(t)}$, agent's intuition $x'^{(t)}$, and the action $x^{(t)}$. Hence, $V^{(t)} = V^{(t)}(s^{(t)}, x'^{(t)})$ and $Q^{(t)} = Q^{(t)}(s^{(t)}, x'^{(t)}, x^{(t)})$. This facilitates the derivation of an insightful result, encoded as theorem (C.3.2) below.

Theorem C.3.2. *Given an MDPUC instance $M\langle\gamma, U, X, Y, S, F, P(u)\rangle$, let $\pi^*_{exp} = \arg\max_{\pi\in F_{exp}} V^\pi(s^{(t)})$ and $\pi^*_{ctf} = \arg\max_{\pi\in F_{ctf}} V^\pi(s^{(t)}, x'^{(t)})$. For any state $s^{(t)}$, the following statement holds:*

$$V^{\pi^*_{exp}}(s^{(t)}) \leq V^{\pi^*_{ctf}}(s^{(t)}).$$

In essence, contemplating counterfactual quantities (intent) never deteriorates performance.

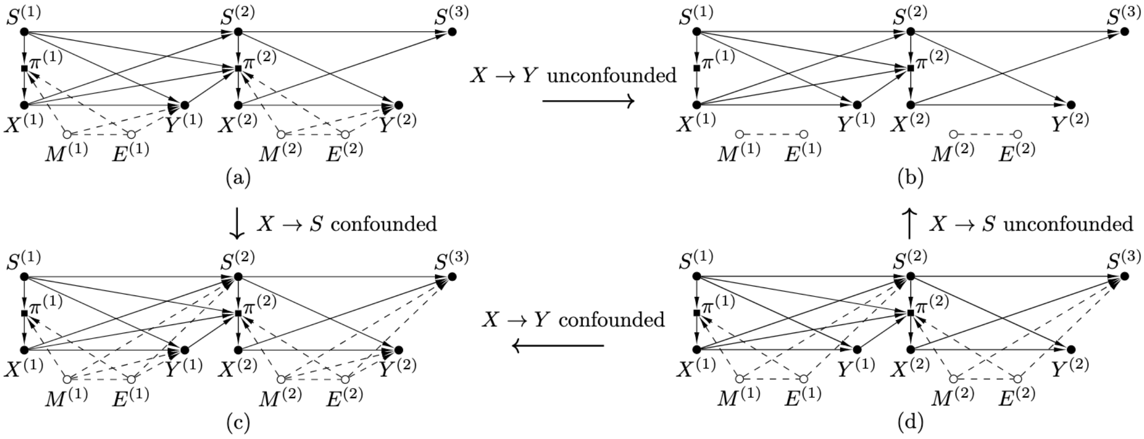


Figure C.1: (a) Depicts an MDPUC with a confounded pathway from action $x^{(t)}$ to reward $y^{(t)}$. (b) Shows a conventional RL model of MDP without any confounders. (c) Illustrates an MDPUC with a confounded route from action $x^{(t)}$ to reward $y^{(t)}$, and also from action to subsequent state $x^{(t)} \to s^{(t+1)}$. (d) Represents an MDPUC where only the action to subsequent state path $x^{(t)} \to s^{(t+1)}$ is confounded. Modified from [139], with a notation variation: X for actions and Y for rewards.

Figure C.1 depicts various configurations of MDPUCs, illustrating how unobserved confounders can affect diverse pathways in decision-making processes [139].

C.4 Learning from Agent Intent

Theorem C.4.1 (Estimating the ETT [109]). *The empirical estimation of the effect of treatment on the treated (ETT) is viable for any number of choices of actions when* *agents base their decisions on a specific intent $I = i$ and then evaluate the response Y to their final action $X = a$.*

Proof. Representing the ETT counterfactual as $\mathbb{E}[Y_{X=a} \mid X = i]$ and utilising the law of total probability alongside the conditional independence $Y_x \perp\!\!\!\perp X \mid I$, we derive:

$$\begin{aligned}\mathbb{E}[Y_{X=a} \mid X = i] &= \sum_{i'} \mathbb{E}[Y_{X=a} \mid X = i, I = i']P(I = i' \mid X = i) \\ &= \sum_{i'} \mathbb{E}[Y_{X=a} \mid I = i']P(I = i' \mid X = i)\end{aligned}$$

Given that $I_x = I$ in $G_{\overline{X}}$ (graph minus edges into X), which implies $(I \perp\!\!\!\perp X)_{G_{\overline{X}}}$, it follows that:

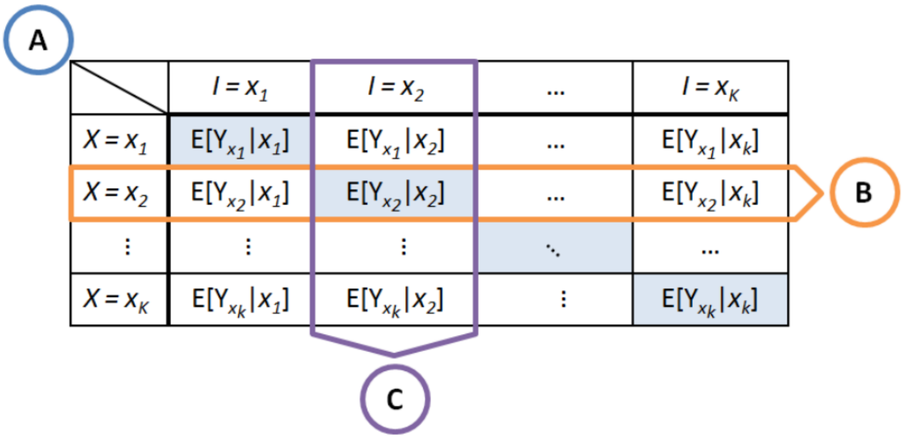
$$\mathbb{E}[Y_{X=a} \mid X = i] = \sum_{i'} \mathbb{E}[Y \mid do(X = a), I = i']P(I = i' \mid X = i)$$

This expresses the counterfactual quantity as an interventional one. Observationally, the intent aligns with the outcome, allowing us to represent this as an indicator function, leading to the conclusion.

$$\mathbb{E}[Y_{X=a} \mid X = i] = \mathbb{E}[Y \mid do(X = a), I = i]$$

□

Forney et al. [109] utilises this theorem to develop heuristics for learning counterfactuals from both experimental and observational data, even when the data might be noisy.



	$I = x_1$	$I = x_2$	...	$I = x_K$
$X = x_1$	$E[Y_{x_1} \mid x_1]$	$E[Y_{x_1} \mid x_2]$	...	$E[Y_{x_1} \mid x_k]$
$X = x_2$	$E[Y_{x_2} \mid x_1]$	$E[Y_{x_2} \mid x_2]$	...	$E[Y_{x_2} \mid x_k]$
⋮	⋮	⋮	⋱	...
$X = x_K$	$E[Y_{x_k} \mid x_1]$	$E[Y_{x_k} \mid x_2]$	⋮	$E[Y_{x_k} \mid x_k]$

Figure C.2: This diagram presents various counterfactual scenarios for distinct actions and their intents. The diagonal line represents the counterfactuals that actually transpired. Utilising known counterfactuals, we can infer about other potential counterfactual outcomes. (B) symbolises learning across different intents, and (C) symbolises learning across different actions. Adapted from [109].

1. **Cross-Intent Learning:** Referencing Equation 4.2 and analysing each arm independently, we establish a set of equations reflecting the outcomes based on varying

intents. For determining $\mathbb{E}[Y_{x_r} \mid x_w]$, insights into *alternate* intent conditions are utilised:

$$\mathbb{E}[Y_{x_r} \mid x_w] = \left[\mathbb{E}[Y_{x_r}] - \sum_{i \neq w}^{K} \mathbb{E}[Y_{x_r} \mid x_i] P(x_i)\right] / P(x_w).$$

2. **Cross-Arm Learning:** In a manner similar to the first point, knowledge about two distinct arms under an identical intent aids in understanding a third arm within the same intent context. This leads to the following expressions:

$$\begin{aligned} P\left(x_w\right) &= \frac{\mathbb{E}\left[Y_{x_r}\right] - \sum_{i \neq w}^{K} \mathbb{E}\left[Y_{x_r} \mid x_i\right] P\left(x_i\right)}{\mathbb{E}\left[Y_{x_r} \mid x_w\right]} \\ &= \frac{\mathbb{E}\left[Y_{x_s}\right] - \sum_{i \neq w}^{K} \mathbb{E}\left[Y_{x_s} \mid x_i\right] P\left(x_i\right)}{\mathbb{E}\left[Y_{x_s} \mid x_w\right]} \end{aligned}$$

By integrating these outcomes, we derive an expression for $\mathbb{E}[Y_{x_r} \mid x_w]$ as follows:

$$\mathbb{E}\left[Y_{x_r} \mid x_w\right] = \frac{\left[\mathbb{E}\left[Y_{x_r}\right] - \sum_{i \neq w}^{K} \mathbb{E}\left[Y_{x_r} \mid x_i\right] P\left(x_i\right)\right] \mathbb{E}\left[Y_{x_s} \mid x_w\right]}{\mathbb{E}\left[Y_{x_s}\right] - \sum_{i \neq w}^{K} \mathbb{E}\left[Y_{x_s} \mid x_i\right] P\left(x_i\right)}. \tag{C.1}$$

However, this estimate is sensitive to sample noise. A pooling method that uses an inverse-variance-weighted average can address reward variance:

$$\mathbb{E}_{XArm}\left[Y_{x_r} \mid x_w\right] = \frac{\sum_{i \neq r}^{K} h_{XArm}\left(x_r, x_w, x_i\right) / \sigma_{x_i, x_w}^2}{\sum_{i \neq r}^{K} 1 / \sigma_{x_i, x_w}^2},$$

where $h_{XArm}\left(x_r, x_w, x_s\right)$ applies equation (C.1), and $\sigma_{x_i,i}^2$ denotes the reward variance for arm x under intent i.

3. **Combined Approach:** Accumulating estimates through execution and integrating both *cross-intent* and *cross-arm* learning methodologies, we deduce a comprehensive approach:

$$\mathbb{E}_{\text{combo}}\left[Y_{x_r} \mid x_w\right] = \frac{\alpha}{\beta}, \quad \text{where}$$

$$\begin{aligned} &\alpha = \mathbb{E}_{\text{samp}}\left[Y_{x_r} \mid x_w\right] / \sigma_{x_r, x_w}^2 + \mathbb{E}_{\text{XInt}}\left[Y_{x_r} \mid x_w\right] / \sigma_{\text{XInt}}^2 + \mathbb{E}_{\text{XArm}}\left[Y_{x_r} \mid x_w\right] / \sigma_{XArm}^2 \\ &\beta = 1 / \sigma_{x_r, x_w}^2 + 1 / \sigma_{\text{XInt}}^2 + 1 / \sigma_{XArm}^2 \end{aligned}$$

C.4.1 Additional Transportability Theory

Definition C.4.1 (mz-Transportability [145])**.** *Consider a set of selection diagrams, $\mathcal{D} = \{D^{(1)}, \ldots, D^{(n)}\}$, originating from source domains $\Pi = \{\pi_1, \ldots, \pi_n\}$ and converging on target domain π^*. Let $\boldsymbol{Z}_i$ denote the variables where experiments can be performed in domain π_i. The causal effect R is deemed mz-transportable from Π to π^* within $\mathcal{D}$ if it can be exclusively computed from the collective observational and interventional distributions.*

The graphic depiction of mz-transportability has an algebraic counterpart derived from the *do*-calculus.

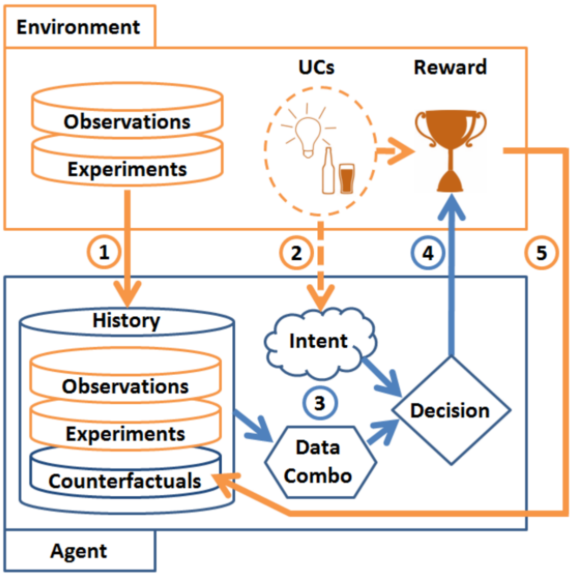


Figure C.3: This illustration outlines the method of integrating data through counterfactual reasoning as discussed in this section. It shows how an agent combines both interventional and observational history to determine counterfactual outcomes. The agent's decision-making, incorporating intended actions, is informed by both counterfactual considerations and intent awareness, taking into account unobserved confounders and maximising the use of available data. Image sourced from [109].

Theorem C.4.2. *Assuming previous definitions, the effect* $R = P^*(\boldsymbol{y} \mid do(x))$ *is mz-transportable from* Π *to* π^* *if the expression* $P(\boldsymbol{y} \mid do(x), \boldsymbol{S}_1, \ldots, \boldsymbol{S}_n)$ *adheres to the do-calculus rules, such that:*

1. *The do-operators, relevant to subsets of* I_z^i*, exclude any* $\boldsymbol{S}_i$*-variables.*

2. *Do-operators are exclusively connected to subsets of* I_z^i*.*

In essence, this theorem reaffirms the completeness of the *do*-calculus for pinpointing transport formulas. For a comprehensive understanding, including an in-depth dive into the established algorithm for deriving these formulas, readers are encouraged to consult the primary source [145].

C.4.2 Causal Structure Learning Theory

Kocaoglu et al. [149] presented a significant enhancement over previous methodologies by offering an algorithm capable of discerning any causal graph. Additionally, it ascertains both the presence and position of latent variables utilising $\mathcal{O}(d\log(n) + l)$ interventions. Here, d signifies the highest node degree, while l represents the most extended directed path in the causal graph. The authors also put forth a probabilistic approach that can learn the observable graph, inclusive of all latent variables, with $\mathcal{O}(d\log^2(n) + d^2\log(n))$ interventions, with high probability. Due to its promising nature, we opted for an in-depth examination and elaboration of this theoretical framework. The authors have partitioned the task of understanding the observable graph and latent variables into three specific stages:

1. First, they introduce a methodology to ascertain the *transitive closure* of the

observable graph.

2. This obtained transitive closure is then *reduced* to spotlight a subset of the edges inherent in the foundational causal graph.

3. Employing conditional independence tests, they proceed to reveal latent variables.

Delving deeper, the authors have demonstrated that *separating systems* can facilitate the formulation of sequences for pairwise conditional independence tests. The aim here is to unveil the transitive closure of the observable causal graph, essentially determining causal pathways by identifying which variables are dependent on others. To articulate this notion rigorously, the post-interventional causal graph concept is invoked. It represents the causal graph G, albeit devoid of edges directed towards intervened variables. A vital consideration here is faithfulness, ensuring causal connections materialise solely due to d-separation, implying the nonexistence of perfectly counterbalancing relations that masquerade as non-causal (see Section 2.11.2).

The formalised conditional independence test is as follows:

Lemma C.4.1 (Pairwise Conditional Independence Test [149])**.** *For a causal graph with latent variables denoted as D_l and an intervention set $S \subset V$ of observable variables, the post-interventional faithfulness presumption stipulates that for any pair $X_i \in S, X_j \in V \setminus S$, $(X_i \not\perp\!\!\!\perp X_j)_{D_l[S]}$ if and only if X_i is an ancestor of X_j in the post-interventional graph $D[S]$.*

This lemma presents methodology to discern ancestry for any variable pair, (X_i, X_j). Yet, its scope is limited. To illustrate, consider $X_i \to X_k \to X_j$ with $X_i, X_k \in S$ and $X_j \notin S$. The authors address this limitation by suggesting a sequence of interventions steered by a separating system, culminating in the discovery of the accurate causal graph through the identification of the transitive closure.

Definition C.4.2 ((m, n) Strongly Separating System [149])**.** *An (m, n) strongly separating system encompasses a collection of subsets $\{S_1, S_2, \ldots, S_m\}$ of the foundational set $[n]$. For any pair of nodes i and j, a set S exists within this collection such that $i \in S, j \notin S$ and vice versa in another set S'.*

The utility of the aforementioned definition is further underscored by findings in [149]. A strong separating system is invariably present on the ground set $[n]$ provided $m \leq 2\lceil \log n \rceil$. This paves the way for the introduction of a deterministic algorithm tailored for deciphering the observable causal graph D derived from ancestral connections. This algorithm requires $2\lceil \log n \rceil$ interventions coupled with conditional independence tests. The key insight is that when the intervention set encapsulates all ancestors of X_i, the only variables showcasing dependence with X_i in the post-interventional cohort are its direct parents, denoted as Pa_i.

Let's consider r as the most extended directed pathway of D_{tc}. Utilising the partial order, $<_{D_{tc}}$, on the vertex ensemble V, it's feasible to derive a distinct partitioning of vertices expressed as $\{T_i \mid i \in [r+1]\}$, where the relationship $T_i <_{tc} T_j$ holds for all $i < j$. Every node within i therefore constitutes a cohort of mutually non-comparable components, symbolising the set of nodes at the i^{th} layer in the transitive closure graph

D_{tc}. If we articulate $\mathcal{T}_i = \cup_{k=1}^{i-1} T_k$, then it's evident that $Pa_i \subset T_i$ – an observation leveraged in the deterministic algorithm proposed by the authors.

Perhaps the standout element of this study is the introduction of a randomised algorithm. The operational strategy is rooted in the recurrent deployment of the ancestor graph learning algorithm to decipher the observable graph. This modus operandi leans heavily on transitive reduction:

Definition C.4.3 (Transitive Reduction)**.** *For a given directed acyclic graph, $D = (V, E)$, with its transitive closure expressed as D_{tc}, the $Tr(D) = (V, E_r)$ represents a directed acyclic graph, characterised by the minimal edge count, such that its transitive closure mirrors D_{tc}.*

Such a transitive reduction is both straightforward and efficient, setting the stage for a repetitive mechanism designed to uncover causal correlations.