Causal Multi-Agent Reinforcement Learning: Review and Open Problems

St John Grimbly[*]
University of Cape Town

Jonathan Shock
University of Cape Town

Arnu Pretorius
InstaDeep

Abstract

This paper serves to introduce the reader to the field of multi-agent reinforcement learning (MARL) and its intersection with methods from the study of causality. We highlight key challenges in MARL and discuss these in the context of how causal methods may assist in tackling them. We promote moving toward a 'causality first' perspective on MARL. Specifically, we argue that causality can offer improved safety, interpretability, and robustness, while also providing strong theoretical guarantees for emergent behaviour. We discuss potential solutions for common challenges, and use this context to motivate future research directions.

1 Introduction

In recent years there has been increasing interest in the possibilities offered to machine learning through a better understanding of causality. Pearl (2019) argues that explicit causal modelling in AI is crucial for achieving general intelligence. The *ladder of causation* is a meta-model which expresses how different types of interaction with a data-generating system (agent-environment interface) limits the types of reasoning an agent can perform. In this sense, reinforcement learning (RL) is seen as a more powerful approach than conventional machine learning methods where purely observational data are used for regression or classification. Despite being able to perform interventions (i.e. take actions) in the environment, without a formal causal model, RL agents lack the ability to explicitly reason counterfactually. Related arguments have prompted interest in the possibility of bridging RL with methods from causality by reformulating RL models/paradigms. This has been attempted with specific tasks in mind, such as off-policy learning (Buesing et al., 2018), data-fusion (Bareinboim and Pearl, 2016; Gasse et al., 2021; Forney et al., 2017), and counterfactual reasoning (Forney and Bareinboim, 2019). This paper considers extending causal RL methods to the multi-agent case, where additional complexities arise due to interacting agents. Further, it posits that causal tools offer appropriate properties for solving some of these challenges.

Reinforcement Learning. The RL problem concerns that of how to map states to actions so as to maximise a numeric reward signal (Sutton and Barto, 2005; Bertsekas, 2012; Levine, 2019). An RL agent attempts to learn to select an optimal sequence of actions by trial-and-error. This reward signal may be partially observed, noisy, or confounded by any number of factors. The difficulty of the learning problem can be exacerbated by stochasticity or non-stationarities in the environment. A useful model for sequential decision making scenarios is defined by the Markov Decision Process (MDP) (Bellman, 1954).

Definition 1.1 (Markov Decision Process (MDP)) *A Markov decision process (MDP) is a stochastic process in which rewards are obtained when transitioning to a new state, dependent on the previous state and the action selected. Formally, it is a 5-tuple (S, A, T, R, γ) of states S, actions A, transition probability function $T = P(s' \mid s, a)$, rewards R, and a discount factor γ.*

---

[*]Work done during an internship at InstaDeep Ltd. Correspondence: me@stjohngrimbly.com

The Markov property implies that the history of how the agent arrived at the current state is irrelevant. This means an optimal policy should prescribe the same action in a particular state regardless of the state-action trajectory. In some problems, agents have limited access to full state information. This is commonly modelled as a Partially-Observable MDP (POMDP) (Åström, 1964). Naturally, the Markov assumption does not hold in general. Consider the *dynamic treatment regime* (DTR) where an individual receives a treatment plan (policy) based on their medical history and unique characteristics (Murphy, 2003; Liu et al., 2019). In a DTR, a doctor should make use of all available information to optimise long term patient outcomes. DTRs are made more difficult by the presence of latent (unobserved) factors that influence variables a healthcare worker has access to. In the case of patient outcomes, there is no way to know the exact effect an intervention will have on an individual over a long time period due to the complexity of the system. For example, remission due to a multiple-course chemotherapy treatment plan could be dependent on initial treatment (Wang et al., 2012). This relationship could be determined by unknown causal factors, in which case no single state would satisfy the Markovian assumption. In general, there may be information or prior knowledge about variables in a system that forms a model over the assumptions. We now discuss ways to formalise causal reasoning within a system of complex interactions and assumptions.

Causality. Causal inference considers how and when causal conclusions can be drawn from data. Recently, there has been great interest in benefits that derive from being explicit about causal assumptions in a modelling procedure. In general, complex systems of interacting variables can be described with a Structural Causal Model (SCM) (Pearl, 2009; Peters et al., 2017). Such a model describes the causal mechanisms and assumptions present in an arbitrary system. The relationships between variables can be graphically presented in the form a *causal Bayesian network* (CBN), wherein nodes represent variables and directed paths represent causal influence between variables (see Figure 1). The reader should not be be confused by related work in Structural Equation Modelling (SEM) which has roots in causal modelling (Pearl, 2012). Pearl (1998) clarifies this confusion about causal assumptions in SEMs by explaining when such methods are valid for claiming causal outcomes. Peters et al. (2017) formulates the SCM as follows:

Definition 1.2 (Structural Causal Model (SCM)) *A structural causal model $M = (S, P_{U_j})$ is a collection S of d structural assignments $X_j \leftarrow f_j(PA_j, U_j)$, $j = 1, \ldots, d$ where PA_j are the parent nodes of X_j, and P_U is a joint distribution over the product of jointly independent noise variables, U_j. An SCM implies a distribution P with density p over variables X in the causal system.*

The key feature of an SCM, as opposed to an MDP, is that it allows for explicit consideration of counterfactual queries. Consider the "does smoking cause cancer?" debate. Perhaps it is possible that there is some hidden genetic factor that causes both cancer and the desire to smoke. Given that it isn't feasible to perform a randomised control trial, an agent must resort to other means of reasoning about this question. Pearl's *ladder of causation* encapsulates the limitations of different systems of reasoning by separating them into three distinct classes (Pearl and Mackenzie, 2018; Pearl, 2019). Firstly, *seeing* (associations) encapsulates statistical reasoning. The second rung corresponds with *doing* (interventions), which contains state-of-the-art randomised control trials (RCTs) and RL methods (Gonzalez-Soto and Espina, 2019; Bareinboim et al., 2020). More specifically, in RL an intervention corresponds to an experiment or *action*, where an agent *intervenes* on the natural state and *causes* a response (e.g. next state and reward). One rung above knowledge accessible by intervention sits *imagination* (counterfactuals). Such counterfactual queries are inherent to the RL problem where an agent could wonder whether it should have previously taken an alternative action. More simply, this formalism allows for explicit reasoning about each action an agent *does*, *can*, and *could* have taken. The nuances of this 'ladder' are formalised in Pearl's Causal Hierarchy (Bareinboim et al., 2020), which establishes the containment relation between different types of interaction with the data-generating processes in a causal model. To be explicit, we provide definitions for interventions and counterfactuals in an SCM.

Definition 1.3 (Intervention) *An intervention I in an SCM M entails changing some set of structural assignments in M with a new set of structural assignments. Assume the replacement is on X_k given by assignment $X_k = \tilde{f}(\tilde{PA_k}, \tilde{U}_k)$, where $\tilde{PA_k}$ are the parents in the underlying new DAG. This change in causal mechanism entails a new interventional distribution, $P_{do(I)}$ and corresponding density $p_{do(I)}$ over the variables of the SCM.*

(a)(i)

$X \rightarrow Y$

(ii) SCM

$X \leftarrow f_x(U_x)$

$Y \leftarrow f_y(X, U_y)$

(b)

$P(Y_x | x', y')$ $\mathcal{L}_3$

$P(Y | do(X))$ $\mathcal{L}_2$

$P(X, Y)$ $\mathcal{L}_1$

Ladder of Causation

(c)

$\mathcal{L}_3$

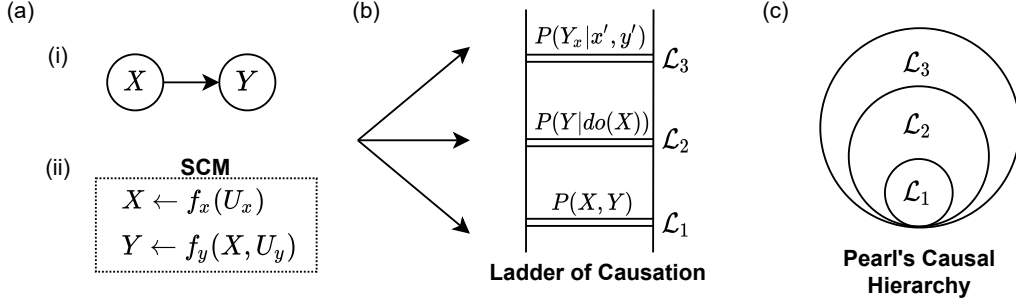$\mathcal{L}_2$

$\mathcal{L}_1$

Pearl's Causal Hierarchy

Figure 1: **(a)(i)** The *causal Bayesian network*, represented by a DAG, depicting causal assumptions about variables and causal directions in a system. **(ii)** An example of a structural causal model (SCM) with two variables of interest corresponding to the DAG in (i). We say X is *exogenous* as it is completely determined by external noise factors, represented by U_x. Y is *endogenous* since it is determined by X and some corresponding noise factor U_y. **(b)** Shows the different types of observed phenomena induced by an SCM. The SCM induces three distinct types of causal quantities, each falling on a separate level of the 'ladder of causation.' Associational (statistical) quantities fall on $\mathcal{L}_1$. Interventions (or experiments) correspond to forcing a variable X to take on value x, thereby removing dependencies on parent variables. We represent this as $do(X = x)$ in Pearl's do-calculus, and these exist on $\mathcal{L}_2$. Finally, counterfactual quantities are indicated by a distribution observed under one set of conditions Y_x, but where we now consider an alternative reality where $X = x'$ such that $Y_x \mid x'$. These exist on $\mathcal{L}_3$. **(c)** Venn diagram showing the containment relation described by Pearl's Causal Hierarchy (Bareinboim et al., 2020), formalising the hierarchy the 'ladder of causation' induces in (b).

Definition 1.4 (Counterfactual) *Consider SCM $M = (S, P_{U_j})$ over nodes X, with observations x. A counterfactual SCM is defined by replacing the distribution of noise variables with the noise associated with the realisation of variables $X = x$. This counterfactual noise term is written as a conditional probability $P_{U|X=x}$.*

With a model established and with knowledge of variables *exogenous* and *endogenous* to the system, methods of *identification* can be used to establish the queries that can be answered within the current model. With this context in mind, we proceed by establishing some key benefits that causality brings to RL, building a context for future research in bridging causality with MARL.

2 Causal Reinforcement Learning

Recent research has looked at tying together methods from causality with the high performance and powerful learning methods of RL, and is referred to as *Causal Reinforcement Learning* (CRL) (Bareinboim, 2020; Li et al., 2021). The structural invariances and performance guarantees gained from explicit causal reasoning, tied together with the learning ability of RL agents, allows for important problems to be considered with a learning-based approach. We now discuss some of the benefits that CRL has established.

Transportability and Data Fusion. A crucial problem in various data-driven machine learning domains is that of mismatched data. When data for learning is collected under mixed policies and environmental conditions, bias makes it difficult to establish rigorous, statistically significant results. Some examples include confounding shift (Landeiro and Culotta, 2018), various types of distribution shift (Levine et al., 2020; Mendonca et al., 2020; Lee et al., 2020), and inductive biases (Hessel et al., 2019). By making assumptions about the data-generating system explicit, and reasoning about the causal relationships between variables of interest, it is possible to establish procedures for algorithmically combining datasets collected under different conditions and policies (Marcellesi, 2015). Bareinboim and Pearl (2016) discuss criteria and techniques for combining datasets curated under different, causally related conditions modelled by an SCM.

A concept related to data fusion is that of *transport*, which asks: when can we feasibly apply a policy learned under one set of conditions when operating under a different set of conditions? Causal

inference can inform as to when and how results can be transported (Pearl and Bareinboim, 2014). A causal model of the underlying system allows for establishing bounds on performance, while also providing techniques for selecting optimal actions (Lee and Bareinboim, 2018, 2020).

Generalised Off-Policy Learning. As is thematic in machine learning in general, RL could benefit immensely from offline and off-policy methods that can learn from existing datasets (Levine et al., 2020). As we discussed in the context of data-fusion and transport, causal inference offers tools for correcting for biases by examining the effect of agent interventions under different policies (Becker, 2016). By reasoning causally, we can combine offline and online modes of learning to effectively improve upon policies - even when we only have observational data (Namkoong et al., 2020). In contrast to conventional supervised and unsupervised methods, the *learning* aspect of RL allows for extracting policies that perform beyond what model fitting of an offline dataset can yield. We can take this a step further by combining methods for offline RL with online environment interaction using a causal model. Zhang and Bareinboim (2017) considers this generalised learning problem in the simplified setting of multi-armed bandits and show improved performance over conventional bandit methods. Multiple papers tackle healthcare problems, such as DTRs (Zhang and Bareinboim, 2019, 2020a; Namkoong et al., 2020), and medical diagnosis (Richens et al., 2020), where guarantees on performance are critical.

Counterfactual Reasoning. Beyond the previous stated benefits, the ability to reason about counterfactual queries is inherently useful. The ability to ask "what if?" is crucial for problems where experiments (taking action) are expensive, safety critical, or simply impossible. Improving the data-efficiency of learning algorithms will require making full use of available information for reasoning. Related to counterfactual reasoning, the idea of imagination within a world model has been studied in the context of RL (Ha and Schmidhuber, 2018). Making such a model causal could provide interesting opportunities for *counterfactual imagination*.

Counterfactual reasoning has motivated new paradigms for MDP-like sequential decision making problems. The MDP with Unobserved Confounders (Bareinboim et al., 2015; Zhang and Bareinboim, 2016), models the setting where state, action, and reward variables are confounded by some latent external factors, potentially causing dependence which can harm learning performance. Further work has considered how consideration for *intended* actions (i.e. the action that would have been taken) can reveal important information about confounders present in the system (Zhang and Bareinboim, 2020b). Especially relevant for RL, a counterfactual approach is also useful for directing exploration towards areas of the system that have causal unknowns and for optimising chosen actions (interventions) within the causal model (Lu et al., 2021; Lee and Bareinboim, 2020). These causal approaches are also being applied to imitation learning, where inference within a causal model allows for guarantees on behaviour (Zhang et al., 2020).

Causal Learning. The process of learning causal structure from data is a rich and active area of research (Glymour et al., 2019; Jaber et al., 2020). Most approaches rely on testing for conditional Independence, but considering other factors, such as time (Löwe et al., 2020) and noise (Hoyer et al., 2008), has proven useful. Some common causal discovery methods include the PC and FCI algorithms Spirtes et al. (2000). The problem posed by causal discovery is related to problems posed in RL literature, where model-based approaches often propose learning a model before using it for planning (Moerland et al., 2020). Further, RL methods have been applied for causal structure learning (Zhu et al., 2019). The authors believe there is great opportunity for merging model-based RL and causal discovery methods.

3 The Multi-Agent Case

Multi-Agent Systems (MAS) is an area of research that studies the interaction of intelligent agents (Wooldridge, 2009). As in RL, multi-agent learning is studied under various models, often extending the RL paradigm to multiple agents (Deming et al., 1944; Littman, 1994; Shoham et al., 2003; Gronauer and Diepold, 2021). A key benefit of the multi-agent approach is the decentralisation of the learning task, which may more naturally map to many potential RL applications. While there has been some interest in merging ideas from causality with current single-agent methods, to the best of our knowledge no research explicitly looks at bridging (graphical) causal methods with MARL. Prior work has looked at exploiting counterfactual knowledge for the multi-agent cooperative setting. For example, Foerster et al. (2018) improve performance on tasks where credit assignment

is challenging by using counterfactual information. Vanneste et al. (2020) also consider the credit assignment problem, improving upon previous results by communicating using counterfactuals. Jaques et al. (2019) propose using causality for intrinsic motivation via social influence. There has also been broader interest in counterfactuals outside of RL. For example, MAS literature considers counterfactuals in terms of *wonderful life utility* (Wolpert et al., 2000). Related work also looked at *difference rewards* as a way to reduce noise and allow agents to learn the 'consequences of their actions' (Devlin et al., 2014).

Causal models are especially relevant in the context of high performance, deep-learning methods which have only recently used for MARL problems (Rashid et al., 2018; Foerster et al., 2016; Vinyals et al., 2019; Zhang et al., 2018). Such multi-agent paradigms immediately afford researchers opportunity to exploit the nature of agent cooperation and interaction, including imitation and communication. Additionally, multiple agents comes with inherent scalablity benefits due to decentralisation of learning and/or execution (Qu et al., 2020; Kaviani et al., 2021; Pretorius et al., 2021). It is our belief that causal methods are especially well suited for dealing with the additional challenges and complexity in multi-agent problems. These complexities are reflected in the various models one finds in the MARL and game theoretic literature (Littman, 1994; Hansen et al., 2004; Bernstein et al., 2002; Deming et al., 1944; Kovařík et al., 2019; Rădulescu et al., 2019). With that said, most modern MARL papers formulate the MARL problem as some multi-agent extension of the MDP. One such extension is the Stochastic (or Markov) Game, where the system evolves in discrete steps according to the combination of state-actions selected by agents in the environment. We can extend stochastic games to the partially observable case where each agent now only has access to some function of the true state (an observation) (Hansen et al., 2004). In the cooperative joint reward and shared history setting, the Partially Observable Stochastic Game (POSG) simplifies to the *decentralised* POMDP (Dec-POMDP) (Bernstein et al., 2002). We provide Definition 3.1 below to highlight the similarity to the definition of the MDP (Definition 1.1). For the sake of clarity of exposition we focus on this simplified setting for the remainder of this paper. With these models in mind, we proceed by shifting our focus to a 'causality first' perspective for multi-agent problems and discuss how this may advance research in the area.

Definition 3.1 (Dec-POMDP) *A Dec-POMDP is a multi-agent extension of a POMDP. It is a 7-tuple $\langle S, \{A_i\}, T, R, \{\Omega_i\}, O, \gamma \rangle$, where S are states, $\{A_i\}$ is the joint action set, $T = P(s' \mid s, a)$ is the set of conditional transition probabilities between states, R is the reward function, $\{\Omega_i\}$ is the joint observation set, $O(s', a, o) = P(o \mid s', a)$ gives the conditional observation distribution, and $\gamma \in [0, 1]$ is the discount factor.*

The multi-agent causal problem has also been considered from a Bayesian learning perspective. The Bayesian Game (Harsanyi, 1967) is a natural formulation for situations where agents have limited information about the actions that other agents take, as well as the reasons they take those actions - their policies. Gonzalez-Soto et al. (2020) extend this idea by including a causal model about the environment. Agents must then maintain a belief about the causal nature of the system. The Bayesian approach leads to a rational decision making criterion, where agents hold probabilistic beliefs about possible causal models of the environment. The game theoretic approach leads to the concept of a *causal Nash equilibrium*, where cooperative agents jointly optimise their actions. The authors feel this is an unexplored and ripe opportunity for application to the MARL problem.

The difficulty of scaling Dec-POMDPs has prompted methods that exploit local dependencies between different agents and their environment. For example, factored Dec-POMDPs (Oliehoek et al., 2008) extend Dec-POMDPs by including *factors*, defined as $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_{|\mathcal{X}|}\}$ that span the state space of the Dec-POMDP. A state is then an assignment of these factors, $s = \{x_1, \ldots, x_{|\mathcal{X}|}\}$. Conditional independence relations between variables in the factored model can be exploited to reduce the dimensionality of the learning problem. Oliehoek et al. (2008) originally suggested using a dynamic Bayesian Network (DBN) to model causal influence in the transition and observation models. The ability for the factored approach to build up an interaction model for the system means it is causally aware at the level of interventions $\mathcal{L}_2$. For the full realisation of causal RL benefits, we really want agent models to be capable of counterfactual reasoning.

Multi-Agent Causal Models (MACM) were proposed in the context of extending causal models to the decentralised multi-agent case, where agents share an environment and have access to private and/or public variables of interest (Maes et al., 2007). Meganck et al. (2005) use MACMs for distributed structure learning, showing that it can be effective in multi-agent causal reasoning problems. The
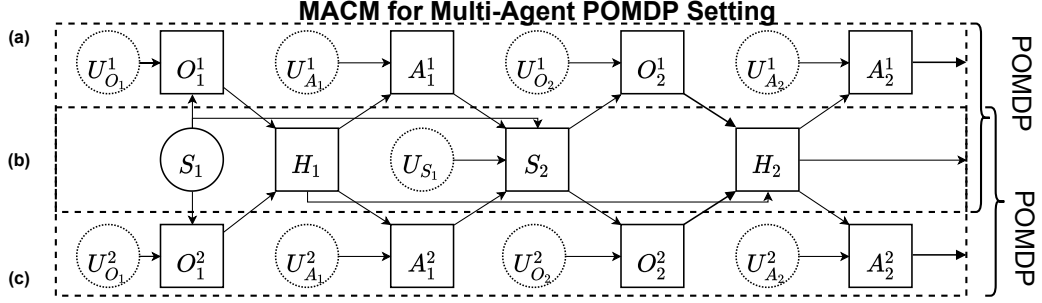
Figure 2: Diagram representing a two-agent sequential decision making system viewed through the lens of an MACM. Squares are used to represent endogenous variables and their associated structural causal assignments. Circles represent exogenous variables, with noise variables highlighted by dotted lines. (a) and (c) can be viewed as individual POMDPs of the agents represented as a Bayesian causal network with noise variables explicitly represented. S represents the global state of the system, O represents observations of the individual agents, H depicts the shared histories (potentially previous observations, actions, and/or rewards depending on problem setting), and A represents actions selected by the individual agents. Agent variables are indicated by superscripts. Noise variables are represented by U. (b) highlights overlap in the POMDPs. The system is an MACM, but we note MACMs can be similarly used to model more general MARL models.

distinguishing factor in MACMs, as opposed to MARL models, is a lack of explicit formalisation of an RL-like learning process. This difference parallels the difference in SCMs and single-agent RL models (see Section 2). This motivates considering reframing common MARL models in terms of MACMs, perhaps allowing for methods that can reason about counterfactual trajectories.

Definition 3.2 (Multi-Agent Causal Models) *A multi-agent causal model (MACM) is a set of n agents, each having access to a semi-Markovian model M_i defined as $M_i = \langle V_{M_i}), G_{M_i}), P(V_{M_i}), K_{M_i} \rangle, \quad i \in \{1, \dots, n\}$. Here V_{M_i} represents the model variables agent i has access to, while G_{M_i} is the causal DAG over V_{M_i}. $P(V_{M_i})$ is the joint distribution over V_{M_i}. Finally, K_{M_i} indicates which variables are shared between agent i and other agents j.*

4 Towards Causal Multi-Agent RL

In this section we examine a way to frame a Dec-POMDP as a multi-agent causal model. We consider the simplified two-agent setting where both agents have limited access to knowledge of the true environment state in the form of observations, but we note that the MACM can be used in a similar way to model more general cooperative scenarios. We discuss one such example where observations are only shared between subsets of agents (see Figure 3).

Returning to the two-agent setting, by modelling state-observation-action trajectories as structural assignments in a causal model we can form two independent SCMs. This means we can be explicit about how the models of interacting agents are related in terms of the components that are shared amongst them. We note that agents share access to the environment and experience the same global underlying state, while also sharing access to the history of state-action trajectories of all agents. In this way, we have a Dec-POMDP modelled as a MACM. We note that although this does not differ substantially from the SCM view of the Dec-POMDP, an MACM formulation handles the sharing of agent variables explicitly whereas an SCM does not. This is useful, for example, in modelling more general scenarios where agent dynamics differ from the Dec-POMDP formulation. Consider the more general scenario where agents must cooperate to achieve optimal outcomes without a shared reward. In this sense, MACMs provide a 'causal wrapper' for general classes of multi-agent models. With this said, we maintain focus on the Dec-POMDP example as it is a foundational model in MARL and is useful for exposition.

To formalise modelling a Dec-POMDP as an MACM, we consider all conditional distributions in the graphical model as being deterministic functions with associated independent noise U (see Figure 2). This includes the transition functions such that a state is causally determined by the previous state, actions and relevant noise terms, $S_{t+1} = f_{st}(S_t, A_t^1, A_t^2, U_{S_t})$. Here S_t represents the environment

state at time-step t, A_t^1 represents actions taken by agent 1 at time-step t, and U_{S_t} is the noise associated with the determination of the underlying state.

Traffic Light Control Example. Consider the four interlinked intersections shown in Figure 3. The goal here is to minimise the total traffic experienced by drivers without full state information. For scaling reasons, we would like to have a decentralised system where each intersection controls the traffic lights given only information about the traffic level on adjacent roads. Modelling this scenario as a Dec-POMDP is possible but is perhaps more efficiently represented in a factored manner. Similarly to the factored Dec-POMDP approach, the decentralised nature of MACMs means we assign each agent an SCM representing its causal model. Crucially, we can explicitly model the shared variables between the neighbours (intersections denoted by I_i) in an efficient manner. For example, I_1 shares traffic observations with both the I_2 and I_3. However, I_2 and I_3 share no observations. The main benefits of this causal modelling approach is that it opens the door for methods that rely on the causal assumptions applied in SCM (and by extension MACM) modelling. We now discuss some specific potential benefits we see emerging in future research in causal MARL.

Tackling Non-stationarity. Individual agents face a moving target problem when other agents simultaneously adapt their policies in response to observed outcomes. This *non-stationarity* arises from limited information about other-agent policies, which is fundamental to the multi-agent, decentralised paradigms (Papoudakis et al., 2019). Common approaches to tackling this include centralising the training procedure (Sharma et al., 2021), accounting for other agents (Raileanu et al., 2018), and applying communication protocols (see section on *Knowledge Sharing* below). Fundamentally, non-stationarity is a source of uncertainty for individual agents. Causal inference offers tools for modelling such uncertain scenarios, especially where agents have structural knowledge of *how* uncertainties are arising (Lee and Bareinboim, 2020; Jesson et al., 2021). Related work has been considering this problem in the Dec-POMDP setting as 'multi-agent beliefs' (see p.59 Oliehoek and Amato, 2016). For example, an agent may know that there is another agent in its vicinity, and it can therefore attribute (some of) the noise in its observations to the unknown behaviour policy of those agents.

Knowledge Sharing. An effective way to increase cooperation in multi-agent systems is by using communication methods (Foerster et al., 2016; Jaques et al., 2019; Das et al., 2019; Kim et al., 2021). Such communication methods are designed to boost learning performance by exploiting existing knowledge encoded in other-agent networks. These communication methods are often less natural than real-world emergent behaviour displayed by human communication, and there is room for research for emergent human intelligible and interpretable communication. One can imagine that communicating causal relationships could greatly boost learning efficiency in decentralised tasks, similar to how Vanneste et al. (2020) improve credit assignment using counterfactual knowledge. This is potentially a fruitful direction for multi-agent, causal research.

Another promising method of knowledge sharing is by applying transfer learning and imitation learning. In these settings, a more knowledgeable agent shares or teaches acquired knowledge to
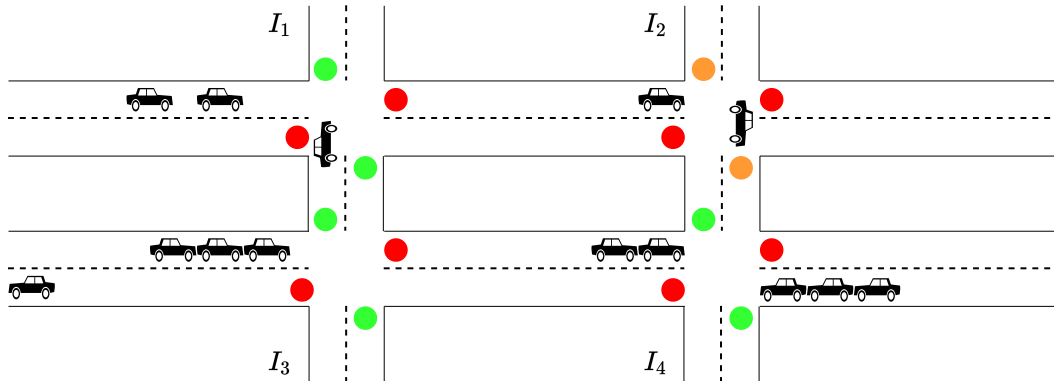

Figure 3: Figure showing system of four intersections, $\{I_1, \ldots, I_4\}$, each treated as an agent in the *traffic light control example* of Section 4. Agents try to minimise global traffic by controlling traffic lights given only measurement of traffic on adjacent roads. Best viewed in colour.

another agent (Pan and Yang, 2010). An exciting CRL research direction is that of causal imitation learning where a causal model of the environment can be used to determine when imitation is feasible, and what performance we can expect. Reasoning causally has shown to aid in determining what variables are important for the imitation procedure (Zhang et al., 2020). Tying these ideas together with (causal) transfer learning and data-fusion opens room for exciting research avenues. For example, teacher-student learning (Zimmer et al., 2014; Omidshafiei et al., 2019; Ilhan et al., 2019), emergent behaviour (Wang et al., 2020), and general cooperative learning could benefit from such ideas.

Decentralised Reasoning. Embedding a causal model in the multi-agent framework opens opportunity for decentralised reasoning in safety critical applications such as healthcare and robotic applications. For example, consider how healthcare practitioners can pool resources (Rieke et al., 2020) and optimise their healthcare policies aided by a safe and interactive MARL systems (Holzinger, 2016). Bridged with other benefits of explicit causal methods already discussed, this presents a promising front for future research. Other topics of interest include multi-agent modelling of market dynamics (Lussange et al., 2020), multi-agent bidding in advertising (Jin et al., 2018), and traffic system control (K.J. et al., 2014).

Generalised Off-Policy MARL. The primary challenge with offline RL is that we are limited to observational data. Viewed from another angle, this problem comes down to learning structure from data collected under mixed policies and is subject to distribution shift (see Section 2). Explicitly modelling the assumptions and uncertainties can aid in bounding the uncertainties in the learning process (Rosenbaum and Rubin, 1983; MacLehose et al., 2005). Extracting maximal information from data is of primary interest in the causal learning literature, and RL researchers interested in this overlap will only benefit as new results emerge. The distribution shift problem is especially exacerbated in the multi-agent problem, where multiple agents with hidden decision policies are influencing the response of the environment (Yang et al., 2021; Levine et al., 2020; Papoudakis et al., 2019). Beyond the fully offline case, many applications could benefit by improving policies with existing offline datasets. This differs from conventional off-policy learning because we lack knowledge of how data was collected. MARL extends the difficulty of this generalised setting by involving more decision agents. We believe this is an interesting research direction.

Model Learning. As in model-based (MA)RL, it may be of interest to first learn a causal model and then use it for identification and planning. Recent causality literature usually frames this as *causal discovery*. These methods have been applied in practice, yielding some impressive results (Hoyer et al., 2008; Zhu et al., 2019; Glymour et al., 2019). Structure learning of multi-agent models has been considered in Meganck et al. (2005). This presents interesting opportunities for research in MARL where, for example, MARL could be used as a heuristic for learning direction of causality by rewarding an agent for correctly choosing actions so as to increase knowledge useful for structure learning.

5 Discussion

In this paper we have introduced selected ideas from causal inference and causal RL with the intention of promoting interest in applying similar methods for MARL. We have discussed a broad range of topics where causality appears to provide interesting avenues for research, including how explicit modelling of causal assumptions can bridge mixed modes of data collection and tackle problems in offline MARL. The authors would like to point out that causal modelling approaches are not limited to graphical model literature, though we have mostly focused on these here as graphical methods are well suited to current approaches to computational sciences in addition to being easy to understand.

There exists extensive literature on *potential outcomes* and *propensity scores* (Rubin, 2005; Rosenbaum and Rubin, 1983). *Instrumental variable* (Becker, 2016) approaches are very common in practice, recently extended as a basis for CRL (Li et al., 2021). To maintain perspective on the current state of causal inference literature, we point to a critiques of SCMs (e.g. Galanti et al., 2020). We also make note of relationships between differential equations and causal models, which are investigated in various settings (Mooij et al., 2013; Bongers and Mooij, 2018; Schölkopf, 2019).

References

J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Commun. ACM*, vol. 62, no. 3, p. 54–60, Feb. 2019. [Online]. Available: https://doi.org/10.1145/3241036

L. Buesing, T. Weber, Y. Zwols, S. Racaniere, A. Guez, J.-B. Lespiau, and N. Heess, "Woulda, coulda, shoulda: Counterfactually-guided policy search," *arXiv preprint arXiv:1811.06272*, 2018.

E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 7345 – 7352, 2016.

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

A. Forney, J. Pearl, and E. Bareinboim, "Counterfactual data-fusion for online reinforcement learners," 2017.

A. Forney and E. Bareinboim, "Counterfactual randomization: Rescuing experimental studies from obscured confounding," 2019.

R. Sutton and A. Barto, "Reinforcement learning: An introduction," *IEEE Transactions on Neural Networks*, vol. 16, pp. 285–286, 2005.

D. Bertsekas, *Dynamic programming and optimal control: Volume I.*   Athena scientific, 2012, vol. 1.

S. Levine, "Deep rl at berkeley: Cs285," http://rail.eecs.berkeley.edu/deeprlcourse/, 2019.

R. Bellman, "The theory of dynamic programming," *Bulletin of the American Mathematical Society*, vol. 60, pp. 503–515, 1954.

K. Åström, "Optimal control of markov processes with incomplete state information," *Journal of Mathematical Analysis and Applications*, vol. 10, pp. 174–205, 1964.

S. Murphy, "Optimal dynamic treatment regimes," *Journal of The Royal Statistical Society Series B-statistical Methodology*, vol. 65, pp. 331–355, 2003.

N. Liu, Y. Liu, B. Logan, Z. Xu, J. Tang, and Y. Wang, "Learning the dynamic treatment regimes from medical registry data through deep q-network," *Scientific Reports*, vol. 9, 2019.

L. Wang, A. Rotnitzky, X. 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*, vol. 107, pp. 493 – 508, 2012.

J. Pearl, *Causality*.   Cambridge university press, 2009.

J. Peters, D. Janzing, and B. Schölkopf, "Elements of causal inference: Foundations and learning algorithms," 2017.

J. Pearl, "The causal foundations of structural equation modeling," 2012.

——, "Graphs, causality, and structural equation models," *Sociological Methods & Research*, vol. 27, pp. 226 – 284, 1998.

J. Pearl and D. Mackenzie, *The Book of Why*.   New York: Basic Books, 2018.

M. Gonzalez-Soto and F. O. Espina, "Reinforcement learning is not a causal problem," *arXiv preprint arXiv:1908.07617*, 2019.

E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard, "On pearl's hierarchy and the foundations of causal inference," *ACM Special Volume in Honor of Judea Pearl (provisional title)*, vol. 2, no. 3, p. 4, 2020.

E. Bareinboim, "Causal reinforcement learning," 2020, iCML 2020. [Online]. Available: https://icml.cc/virtual/2020/tutorial/5752

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

V. Landeiro and A. Culotta, "Robust text classification under confounding shift," *J. Artif. Intell. Res.*, vol. 63, pp. 391–419, 2018.

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

R. Mendonca, X. Geng, C. Finn, and S. Levine, "Meta-reinforcement learning robust to distributional shift via model identification and experience relabeling," *arXiv preprint arXiv:2006.07178*, 2020.

S. Lee, Y. Seo, K. Lee, P. Abbeel, and J. Shin, "Addressing distribution shift in online reinforcement learning with offline datasets," 2020.

M. Hessel, H. van Hasselt, J. Modayil, and D. Silver, "On inductive biases in deep reinforcement learning," *arXiv preprint arXiv:1907.02908*, 2019.

A. Marcellesi, "External validity: Is there still a problem?" *Philosophy of Science*, vol. 82, pp. 1308 – 1317, 2015.

J. Pearl and E. Bareinboim, "External validity: From do-calculus to transportability across populations," *Statistical Science*, vol. 29, no. 4, Nov 2014. [Online]. Available: http://dx.doi.org/10.1214/14-STS486

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

——, "Characterizing optimal mixed policies: Where to intervene and what to observe," *Advances in neural information processing systems*, vol. 33, 2020.

S. O. Becker, "Using instrumental variables to establish causality," *The IZA World of Labor*, pp. 250–250, 2016.

H. Namkoong, R. Keramati, S. Yadlowsky, and E. Brunskill, "Off-policy policy evaluation for sequential decisions under unobserved confounding," *arXiv preprint arXiv:2003.05623*, 2020.

J. Zhang and E. Bareinboim, "Transfer learning in multi-armed bandits: A causal approach," pp. 1340–1346, 2017. [Online]. Available: https://doi.org/10.24963/ijcai.2017/186

——, "Near-optimal reinforcement learning in dynamic treatment regimes," 2019.

——, "Designing optimal dynamic treatment regimes: A causal reinforcement learning approach," 2020.

J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature Communications*, vol. 11, 2020.

D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

E. Bareinboim, A. Forney, and J. Pearl, "Bandits with unobserved confounders: A causal approach," 2015.

J. Zhang and E. Bareinboim, "Markov decision processes with unobserved confounders: A causal approach," 2016.

——, "Can humans be out of the loop?" 2020.

Y. Lu, A. Meisami, and A. Tewari, "Causal markov decision processes: Learning good interventions efficiently," *arXiv preprint arXiv:2102.07663*, 2021.

J. Zhang, D. Kumor, and E. Bareinboim, "Causal imitation learning with unobserved confounders," *Advances in neural information processing systems*, vol. 33, 2020.

C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, 2019.

A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim, "Causal discovery from soft interventions with unknown targets: Characterization and learning," *Advances in neural information processing systems*, vol. 33, 2020.

S. Löwe, D. Madras, R. Zemel, and M. Welling, "Amortized causal discovery: Learning to infer causal graphs from time-series data," *arXiv preprint arXiv:2006.10833*, 2020.

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," 2008.

P. Spirtes, C. Glymour, and R. Scheines, "Causation, prediction, and search, second edition," 2000.

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

S. Zhu, I. Ng, and Z. Chen, "Causal discovery with reinforcement learning," *arXiv preprint arXiv:1906.04477*, 2019.

M. Wooldridge, *An introduction to multiagent systems*.   John wiley & sons, 2009.

W. Deming, J. Neumann, and O. Morgenstern, "Theory of games and economic behavior." *Journal of the American Statistical Association*, vol. 40, p. 263, 1944.

M. Littman, "Markov games as a framework for multi-agent reinforcement learning," 1994.

Y. Shoham, R. Powers, and T. Grenager, "Multi-agent reinforcement learning: a critical survey," 2003.

S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, pp. 1–49, 2021.

J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," vol. 32, no. 1, 2018.

S. Vanneste, A. Vanneste, K. Mets, A. Anwar, S. Mercelis, S. Latré, and P. Hellinckx, "Learning to communicate using counterfactual reasoning," *arXiv preprint arXiv:2006.07200*, 2020.

N. Jaques, A. Lazaridou, E. Hughes, Çaglar Gülçehre, P. A. Ortega, D. Strouse, J. Z. Leibo, and N. D. Freitas, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," 2019.

D. H. Wolpert, K. Tumer, and K. Swanson, "Optimal wonderful life utility functions in multi-agent systems," 2000.

S. Devlin, L. Yliniemi, D. Kudenko, and K. Tumer, "Potential-based difference rewards for multiagent reinforcement learning," pp. 165–172, 2014.

T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," pp. 4295–4304, 2018.

J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," 2016.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. Agapiou, M. Jaderberg, A. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. 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*, pp. 1–5, 2019.

K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Fully decentralized multi-agent reinforcement learning with networked agents," 2018.

G. Qu, Y. Lin, A. Wierman, and N. Li, "Scalable multi-agent reinforcement learning for networked systems with average reward," *arXiv preprint arXiv:2006.06626*, 2020.

S. Kaviani, B. Ryu, E. Ahmed, K. A. Larson, A. Le, A. Yahja, and J. H. Kim, "Robust and scalable routing with multi-agent deep reinforcement learning for manets," *arXiv preprint arXiv:2101.03273*, 2021.

A. Pretorius, K. ab Tessera, A. P. Smit, C. Formanek, S. J. Grimbly, K. Eloff, S. Danisa, L. Francis, J. Shock, H. Kamper, W. Brink, H. Engelbrecht, A. Laterre, and K. Beguir, "Mava: a research framework for distributed multi-agent reinforcement learning," 2021.

E. Hansen, D. Bernstein, and S. Zilberstein, "Dynamic programming for partially observable stochastic games," 2004.

D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Mathematics of operations research*, vol. 27, no. 4, pp. 819–840, 2002.

V. Kovařík, M. Schmid, N. Burch, M. Bowling, and V. Lisỳ, "Rethinking formal models of partially observable multiagent decision making," *arXiv preprint arXiv:1906.11110*, 2019.

R. Rădulescu, P. Mannion, D. M. Roijers, and A. Nowé, "Multi-objective multi-agent decision making: a utility-based analysis and survey," *Autonomous Agents and Multi-Agent Systems*, vol. 34, no. 1, Dec 2019. [Online]. Available: http://dx.doi.org/10.1007/s10458-019-09433-x

J. Harsanyi, "Games with incomplete information played by "bayesian" players, i-iii: Part i. the basic model&," *Manag. Sci.*, vol. 14, pp. 159–182, 1967.

M. Gonzalez-Soto, L. Sucar, and H. Escalante, "Causal games and causal nash equilibrium," *Res. Comput. Sci.*, vol. 149, pp. 123–133, 2020.

F. A. Oliehoek, M. T. Spaan, N. Vlassis, and S. Whiteson, "Exploiting locality of interaction in factored dec-pomdps," pp. 517–524, 2008.

S. Maes, S. Meganck, and B. Manderick, "Inference in multi-agent causal models," *Int. J. Approx. Reason.*, vol. 46, pp. 274–299, 2007.

S. Meganck, S. Maes, B. Manderick, and P. Leray, "Distributed learning of multi-agent causal models," pp. 285–288, 2005.

G. Papoudakis, F. Christianos, A. Rahman, and S. V. Albrecht, "Dealing with non-stationarity in multi-agent deep reinforcement learning," *arXiv preprint arXiv:1906.04737*, 2019.

P. Sharma, R. Fernandez, E. G. Zaroukian, M. Dorothy, A. Basak, and D. E. Asher, "Survey of recent multi-agent reinforcement learning algorithms utilizing centralized training," 2021.

R. Raileanu, E. L. Denton, A. D. Szlam, and R. Fergus, "Modeling others using oneself in multi-agent reinforcement learning," 2018.

A. Jesson, S. Mindermann, Y. Gal, and U. Shalit, "Quantifying ignorance in individual-level causal-effect estimates under hidden confounding," *arXiv preprint arXiv:2103.04850*, 2021.

F. A. Oliehoek and C. Amato, *A concise introduction to decentralized POMDPs*. Springer, 2016.

A. Das, T. Gervet, J. Romoff, D. Batra, D. Parikh, M. Rabbat, and J. Pineau, "Tarmac: Targeted multi-agent communication," pp. 1538–1546, 2019.

W. Kim, J. Park, and Y. Sung, "Communication in multi-agent reinforcement learning: Intention sharing," 2021.

S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, 2010.

M. Zimmer, P. Viappiani, and P. Weng, "Teacher-student framework: a reinforcement learning approach," 2014.

S. Omidshafiei, D.-K. Kim, M. Liu, G. Tesauro, M. Riemer, C. Amato, M. Campbell, and J. How, "Learning to teach in cooperative multiagent reinforcement learning," 2019.

E. Ilhan, J. Gow, and D. P. Liebana, "Teaching on a budget in multi-agent deep reinforcement learning," *2019 IEEE Conference on Games (CoG)*, pp. 1–8, 2019.

T. Wang, H. Dong, V. Lesser, and C. Zhang, "Roma: Multi-agent reinforcement learning with emergent roles," *arXiv preprint arXiv:2003.08039*, 2020.

N. Rieke, J. Hancox, W. Li, F. Milletari, H. Roth, S. Albarqouni, S. Bakas, M. Galtier, B. Landman, K. H. Maier-Hein, S. Ourselin, M. J. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. Cardoso, "The future of digital health with federated learning," *NPJ Digital Medicine*, vol. 3, 2020.

A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, pp. 119 – 131, 2016.

J. Lussange, I. Lazarevich, S. Bourgeois-Gironde, S. Palminteri, and B. Gutkin, "Modelling stock markets by multi-agent reinforcement learning," *Computational Economics*, vol. 57, pp. 113–147, 2020.

J. Jin, C. Song, H. Li, K. Gai, J. Wang, and W. 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.

P. K.J., H. K. A.N, and S. Bhatnagar, "Multi-agent reinforcement learning for traffic signal control," pp. 2529–2534, 2014.

P. Rosenbaum and D. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, pp. 41–55, 1983.

R. MacLehose, S. Kaufman, J. Kaufman, and C. Poole, "Bounding causal effects under uncontrolled confounding using counterfactuals," *Epidemiology*, vol. 16, pp. 548–555, 2005.

Y. Yang, X. Ma, C. Li, Z. Zheng, Q. Zhang, G. Huang, J. Yang, and Q. Zhao, "Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning," *arXiv preprint arXiv:2106.03400*, 2021.

D. Rubin, "Causal inference using potential outcomes," *Journal of the American Statistical Association*, vol. 100, pp. 322 – 331, 2005.

T. Galanti, O. Nabati, and L. Wolf, "A critical view of the structural causal model," *arXiv preprint arXiv:2002.10007*, 2020.

J. M. Mooij, D. Janzing, and B. Schölkopf, "From ordinary differential equations to structural causal models: the deterministic case," *arXiv preprint arXiv:1304.7920*, 2013.

S. Bongers and J. M. Mooij, "From random differential equations to structural causal models: The stochastic case," *arXiv preprint arXiv:1803.08784*, 2018.

B. Schölkopf, "Causality for machine learning," *arXiv preprint arXiv:1911.10500*, 2019.