Faithfulness violations due to the determinability of effects in certain classes of discrete and discretisable Markovian causal models

Pouya Babakhani

MSc in Artificial Intelligence
The University of Bath
2024

Faithfulness violations due to the determinability of effects in certain classes of discrete and discretisable Markovian causal models

Faithfulness violations due to the determinability of effects in certain classes of discrete and discretisable Markovian causal models

Submitted by: Pouya Babakhani

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of MSc in Artificial Intelligence in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Faithfulness violations due to the determinability of effects in certain classes of discrete and discretisable Markovian causal models

Abstract

The task of learning about the causal structure of a system from data on a set of system variables is termed causal discovery (Spirtes and Zhang, 2016; Zanga and Stella, 2022). Often, most if not all the data that is available for causal discovery will be observational, as opposed to interventional, meaning that the data will come from the passive observation of the system, as opposed to data obtained from carrying out purposeful manipulations or interventions in the system, e.g. via experimentation. Under such an observational setting, stronger assumptions about the properties of the causal dependencies of a system are needed for causal discovery to be feasible. One such assumption that is commonly made is the causal faithfulness condition (CFC or just FC), roughly stating that variables that are causally "connected" should also be measured to be probabilistically dependent. Since this condition may be violated, theoretical justifications have been made, arguing that scenarios where violations of the condition occur are rare, e.g. obtain with probability 0 over certain parametrisations of spaces of possible systems, in turn legitimising the presumption of the FC (Meek, 1995; Spirtes et al., 2001, p. 41), and by extension, the usage of causal discovery methods that depend on the condition holding. However, in this thesis we critique such arguments by showing how the prevalence of faithfulness violations is very much contingent on the assumptions we make about the systems we are dealing with. We will demonstrate 3 distinct conceivable settings under which we can expect to see faithfulness violations with positive probability. So, our work serves as a critique of the "probability 0" theoretical defences that have been made in support for the FC and, by extension, the reliability of the causal discovery methods that depend upon it.

Faithfulness violations due to the determinability of effects in certain classes of discrete and discretisable Markovian causal models

Contents

List of Figures

Acknowledgements

I would like to thank my wife and family for supporting me. I would also like to thank my supervisor Dr. Paola Bruscoli for her kind and valuable advice, patience and time.

Chapter 1

Introduction and background

1.1 Introduction

One of the surest ways to determine if some event A influences the realisation of another event B, that is to determine if A causes B, is to carry out an experiment where A is induced in an artificial and targeted manner, and then to check whether B obtains accordingly (Tiles, 1993). However, experiments can be expensive, slow, dangerous, unethical or even impossible. As an example, we cannot manipulate distant celestial objects, so how do derive knowledge about how distant bodies function and interact? Armed only with prior knowledge, and limited to passive observation, we may call the task of acquiring more knowledge about how some system works in terms of cause-and-effect phenomena as *causal discovery without interventions* or just causal discovery (Spirtes and Zhang, 2016). What's more, many formalised approaches and algorithms that have been developed in recent decades for automating the task of causal discovery that assume the so called *causal faithfulness condition* (FCC or just FC): the idea being, roughly speaking, that whenever A and B are causally connected in a system, then, as we look at many system instances, there should be a probabilistic dependence between A and B regardless of what is happening in the system (Spirtes et al., 2001, p. 31; Weinberger, 2018). In particular, the only means of rendering causally connected variables A and B probabilistically independent should be by conditioning on *mediating* variables for *every* pathway of causal influence linking A and B. If this can be achieved, then every means for the transmission causal influence from one variable to the other will be *blocked*, and the variable pair should then be probabilistically independent as there is no reason to expect them to be otherwise spontaneously coordinated. What we have just paraphrased is Reichenbach's (1956) influential *common cause principle*, which essentially states that if 2 types of events A and B are statistically dependent, then there must be a causal process in nature connecting them. Either the occurrence of A causes (i.e. increases the chance of) B, B causes A or they share a common cause C. Moreover, controlling for a common cause C should *screen off* A from B, i.e. make them independent. Reichenbach notably argued that his principle obtains as a macroscopic artifact of statistical mechanics, though there are

objections to this view (Hitchcock and Rédei, 2021), but we digress. The essential idea to appreciate is that the FC and the common cause principle jointly suggest that we can identify direct, i.e. unmediated causal dependence between some *A* and *B* through passive observation, because, granted the FC, causally directly linked events should *always* appear to be probabilistically dependent, whereas events that are not directly linked may be observed to be probabilistically independent granted the right conditioning by virtue of the common cause principle. This distinction is the key to many causal discovery methods and therefore, a major concern is that the FC condition can be violated. Indeed, a notable class of systems that appear to feature FC violations are those that maintain steady states against disturbances, e.g. biological systems (Andersen, 2013; Weinberger, 2023). For example, without the right data, we may find that drinking water appears to have little to no influence on the total amount of water in the body (due to the body maintaining fluid balance via processes such as urination). However, such an apparent independence should not prompt the incorrect conclusion that drinking water does not hydrate us, as might be falsely presumed by causal discovery methods that depend on the FC condition. In turn, such methods have been defended by arguing that any such scenarios featuring violations of FC, are in a sense, going to be very rare and far-in-between out of all possible systems and causal mechanisms that are conceivable (Spirtes et al., 2001, p. 41). Indeed, if some system can be described in discrete and finite terns, i.e. if its number of possible distinct states is finite, and granted some additional assumptions that we will review, then the probability of there being causal dependences in the system (e.g. between distinct sub-components of the system) that correspond with violations of FC, i.e. that coincide with probabilistic independences, has be shown to be 0 (Meek, 1995).

Now, the central aim of our thesis is to highlight how such "probability 0" arguments are contingent on assumptions that may violated, as it is possible to conceive of alternative assumptions, which we will detail and justify (in chapter 2), under which the probability of faithfulness violations can be demonstrated to be greater than 0. In particular, we will consider 2 discrete settings and a continuous setting under which FC violations will occur with positive probability due to what we term as *effect determinability*: the phenomenon where a set of conditioning variables constrain a causal variable to the extent that its effect on another variable can be determined. Again, we will justify our settings by arguing that the assumptions that we introduce are plausible, yet our intention is not be to prove, in any sense, that our assumptions are more *probable* in the real world than the assumptions from e.g. Meek's setting; we merely seek to demonstrate that it is precarious to draw any conclusions about the frequency of a phenomenon such as FC violations, when the answer very much depends on the assumptions, i.e. types of systems, we presume we are dealing with.

The remainder of this thesis is structured as follows: the rest of this chapter is dedicated to an extensive discussion on a range of background matters that will assist us in making sense of our theoretical and experimental results in chapter 2, starting with a slightly more formal and philosophical treatment of the subject matter of causality in section 1.2. In sections 1.3 and 1.4, we will, via our own approach, attempt to build a bridge between the more philosophical content of section 1.2 and the more computation and practical aspects of causality, in particular the framework of structural causal models (SCMs). In section 1.5, we will review the theory of Bayesian networks (BNs), and briefly discuss Meek's (1995) result. In 1.6, we will see how SCMs and BNs interlink, paying particular attention to FC violations. Finally, as stated, chapter 2 will present our results, and chapter 3 will provide a brief conclusion.

1.2 Causality

1.2.1 Causal intuition and causality in science and AI

Observing a person putting on a coat at home is likely a good predictor of whether that person will be going outside. Granted a large dataset of pictures showing peoples' faces, the presence of sunglasses is likely a decent predictor of whether it was sunny at the time and place the pictures were taken at. However, does putting on a trench coat *cause* one to be physically transported out of buildings? Does putting on sunglasses *cause* the sun to shine? These rhetorical questions highlight that statistical dependence does not equate with causality.

The nature of causality has been the subject of inquiry for millennia (Falcon, 2023), and its relationship with statistics is an active area of research. Yet it seems that we, as humans, possess a natural faculty for thinking in causal terms, and we may conjecture that there is an evolutionary reason for our capacity for "causal cognition" (Lombard and Gärdenfors, 2017). Consider that for survival, it may not be enough to passively predict the dynamics of our environment; we must also know how to effectively *direct* our surroundings towards beneficial events. For example, we must act in a manner that yields food and prevents us from becoming food. By having causal knowledge about the influence of different actions on environmental variables, and a model of how these different variables themselves are causally interlinked, we can evaluate, by iterating through such a causal model of the world, whether a certain action could have useful downstream effects.

Furthermore, though historically a matter of contention (Russell, 1912), elucidating the causal mechanisms that underly natural phenomena can be argued to encapsulate what scientific discovery is essentially about (Woodward and Ross, 2021). For example, consider a randomized controlled trial for a treatment: such a procedure can be understood as an attempt to estimate the causal influence, if any, that a treatment has on some health outcomes. Taking matters further, some conceive of a future where AI systems independently design experiments to test and improve causal hypotheses (Tigas et al., 2022; Toth et al., 2022; Zhang et al., 2023), and where AI systems may explain their reasoning and decision making via references to chains of cause-and-effects pairs (Pearl and Mackenzie, 2018; Carloni, Berti and Colantonio, 2023). Causal learning may also facilitate generalization and out-of-distribution prediction in machine learning (Peters, Bühlmann and Meinshausen, 2016; Bengio et al., 2019; Schölkopf et al., 2021), as it appears that most of the causal mechanisms of nature are stable (Pearl, 2009, p. 25) in the sense that they vary little, if at all, across different environments, and fittingly, large statistical differences between environments may be explained in terms of only a few tweaked causal mechanisms. Accordingly, models that "factorize" the world in terms of causally valid dependencies/mechanisms may adapt to environmental shifts via only a few parameter adjustments, whereas models that utilize many causally invalid conditional dependencies may struggle, as many such dependencies can collectively shift and require re-estimation. In summary, principles from the study of causality may be valuable for developing generalizable and explainable AI systems capable of independently deriving new knowledge.

1.2.2 Causation in terms of counterfactuals

Having briefly motivated the notion of causality and noted its distinctness from statistical association, we may benefit from a discussion on counterfactual definitions of causation. Note, we distinguish between the terms "causality" and "causation" as follows: causality refers to the general subject of causal relationships, whereas causation is limited to *unconfounded* and direct causal relationships (i.e. not due to a common cause). In turn, if we can thoroughly understand causation, extending our grasp to causality should be relatively straightforward.

There are various lenses through which causation can be analysed, e.g. in terms of regularity, process, probabilistic, interventionistic or counterfactual theories, (Beebee, Hitchcock and Menzies, 2009, pp. 131-263). Yet independent of the framework, we may distinguish between token and type causation. A statement about token causation is a claim about a specific instance of events, for example: "Timmy has a headache today because he was drunk last night". Type causation, in turn, concerns general rules that apply beyond specific instances: "getting drunk causes headaches the next day". We may also distinguish the above forms of *constant* causation from *variable* causation, also called *influence*: "The more Timmy drinks tonight, the worse his headache will be tomorrow" is a statement about token influence (token variable causation), and "the more one drinks, the worse their headache will be the next day" is a statement about type influence (Gallow, 2022). Though causal modelling is, in practice mainly applied to the study of type influence, the philosophical literature on causality focuses on token, singular, or *actual*, causation (and so will this section), as the analysis of token causation can be conceptually simpler and more in-tune with our every-day intuition. What's more, it should be feasible to extrapolate definitions for actual causation to type influence.

A notable counterfactual definition of actual causation was given by David Lewis (1973a):

> (A) Where *c* and *e* are two distinct possible events, *e* causally depends on *c* if and only if, if *c* were to occur *e* would occur; and if *c* were not to occur *e* would not occur.

Lewis elaborates:

> We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well.

What renders definition (A) *counterfactual* is that for a given pair of actual events, the qualification of causation hinges on the truthfulness of a counterfactual dependence of (the occurrence of) the event purported to be the effect on the cause. However, in reality we cannot test what would have been the case if a particular past event that has occurred had instead not occurred, likely rendering causal claims made in the sense above impossible to assess in practice without, at least, an appeal of some form to a notion of regularity in nature. Indeed, the challenge of having to assess counterfactual claims to arrive at causal knowledge has been referred to as the "fundamental problem of causal inference" (Holland, 1986).

Moving on, there are other pressing issues with the Lewisian definition (A) that need to be dealt with. For example, we must clarify what constitutes an *event*. Events have been defined as the contents of specific regions of spacetime (Quine, 1985), yet this view may be too coarse-grained: consider the example of a spinning hot metal ball from Davidson (1969). The ball

may have been spun by a passerby, and its hotness could be due to contact with a hot surface. Now, the heat of the ball may be causing the air near the ball to warm, whereas the spin may have no such effect. Yet, we are unable to distinguish the ball's rotation from its temperature, as both properties coincide over the same spacetime region and so constitute the same event.

Kim (1976) defines an event as a triple of an object, a property, and a time interval. Substituting time intervals with spacetime regions yields an extension of the definition of Quine, yet viewing events as such Kimian triples may still not be fine-grained enough, insofar as it may not be clearly expressible what exact aspect of an event is causally pertinent. Yet, instead of resorting to even more fine-grained causal relata such as *event allomorphs* (Dretske, 1977), we may instead take issue with the arity of definition (A). Being binary, it does not permit any description of its implicit counterfactual. Consider the statement "Smoking a pack a day gave Jimmy lung cancer"; we may elaborate and say that "smoking a pack a day, as opposed to not smoking at all, gave Jimmy lung cancer", which is more informative and helps us avoid misconstruing something like "smoking one pack a day as opposed to smoking two packs a day is what gave Jimmy lung cancer". Contrasts with respect to the effect also help characterize the nature of the causal impact: Consider the statement "Lulu's hunger caused her to eat an apple, as opposed to eating nothing", which makes more sense than "Lulu's hunger caused her to eat an apple, as opposed to eating an orange". For all we know, hunger may be a phenomenon that simply drives one to avoid eating citrus fruits, but stating the right contrast prevents such a misunderstanding. Schaffer (2005) by including contrasts for both the cause and effect, defines causation as the quaternary relation: c rather than c^* caused e rather than e^*. For example: "Bob stealing the bike, rather than paying for it, caused him to be arrested, rather than to be cycling home". We are made to understand that stealing causes arrest, e.g. as opposed to the possession of bicycles in general, and that the alternative to arrest would have been freedom and not, for example, both arrest and corporal punishment.

So far, we have considered defining events as triples of objects, properties and spacetime regions, and token causation as a quaternary relation among events employing counterfactual contrasts. However, issues still abound, yet with this only being a short overview of the topic of counterfactual causation, we conclude with a brief discussion of some such issues.

Firstly, recall that counterfactuals are unactualized possibilities that, by definition, cannot be verified by observation. Yet we must consider how counterfactual claims can, at least in theory, be assessed. Lewis (1973b), utilising *possible world semantics*, ascertains the veracity of a causal dependence if it would be true that in the hypothetical world *closest* to our actual world (in terms of its content of events and its physical laws), with equivalence right up until the moment of the occurrence of the cause, an intervention via a small targeted "miracle" that would only alter or negate the cause would be followed by an alteration or negation of the effect, barring any further miracles, that is, contraventions against the laws of nature. Resorting to a notion of *miracles* suggests that evaluating counterfactuals may not be so straightforward.

Secondly, the study of causality must contend with absences or *non-events*. Consider "the plant died because Timur did not water it". Now, whether non-events such as Timur *not* watering his plant can have any causal impacts is a matter of contention. Armstrong (1999) asserts:

> Omissions and so forth are not part of the real driving force in nature. Every causal situation develops as it does as a result of the presence of positive factors alone.

Indeed, process theories that define causation through criteria regarding the flow of e.g. momentum-energy, information or "marks" (Dowe, 2009) can fail to identify any meaningful physical process connecting what Timur was doing instead of watering his plant (e.g. being abroad on holiday) with the death of his plant. In contrast, non-events seem to be better handled by counterfactual accounts, as one can specify a "non-causation" counterfactually: "Timur going on holiday to Cape Verde for two weeks, *rather than being at home and regularly watering his plant*, caused his plant to die, *rather than to continue living on*."

With that being said, the fortunes of process and counterfactual theories can reverse when counterfactuals fail to make a difference. Consider the following vignette about "pre-emption" due to Hall (2004): Suzy and Billy both decide to throw rocks at a bottle. Suzy throws first (or faster) and her rock smashes the bottle. Right after, Billy's rock flies over the space where the bottle had been standing a moment ago. It appears, Suzy's throw pre-emptively broke the bottle before Billy's rock arrived, and had Suzy not hit the bottle, e.g. by missing, then Billy's rock would have done the job. As a result, we cannot claim "Suzy's throw caused the bottle to break" if we regard causation in a strictly counterfactual sense, because the breaking of the bottle did not counterfactually depend on Suzy's throw. In contrast, process-oriented theories of causality can more readily identify Suzy's throw as the cause of the breakage, as Billy's rock did not make contact and/or transfer any information or energy-like quantity to the bottle.

Moving on, we must also draw a distinction between causation and logical implication. That is, if some event logically implies some other event, it does not follow that the events are causally linked. E.g. the events of putting on a blue-colored jacket and taking a train to Paddington station do not *cause* one to be a bluish person in London. Rather, the latter is logically implied by the former. Lewis (1973a) seeking to preclude cases of logical implication from causation, formulated a requirement on the *distinctness* of events. As a negative example from Menzies and Beebee (2020), consider that writing "Larry" is not a cause of writing an "rr". So, causal dependence between some pair of events may only come into consideration after certifying, in some sense, their semantical/logical distinctness.

In this section, we have only scratched the surface of the philosophical study of causation and limited ourselves to counterfactual token causation, yet the general essence of the counterfactual view should be apparent: causal dependence exists, when under some circumstances (presumably under the given ones), a change *solely* to the cause would suffice to bring about a change in the effect. For more extensive introductions to counterfactual causation, we refers readers to Menzies and Beebee (2020) and Gallow (2022).

1.3 Systems, measurements and causal mechanisms

1.3.1 Conceivable and plausible system events and states

In practice, we aim to study *type influence*, i.e. to understand the enduring causal structure of a particular system, or of any system of a certain type. The study of type influence is facilitated by positing the existence of stable causal *mechanisms* in our systems, which may be interpreted as invariant rules that govern what further events are causally entailed as effects from a given collection of actual events (Waldmann, 2017, pp. 127-146). We will soon see in section 1.3.2

how counterfactual causation can be used to develop the notion of a causal mechanism.

For the purposes of our thesis, let a system simply be a bounded region of spacetime, and let a system state s with respect to a time or system instance t be characterised as the collection of events that obtain in the system at instance t, out of a collection $\mathbf{\Omega}$ of *conceivable* events that could obtain within the bounds of the system. Iff event $e \in \mathbf{\Omega}$ obtains in state s, we state $e \in s$. Indeed, a state s is considered *fully specified* iff $\forall e \in \mathbf{\Omega}$, we can answer whether $e \in s$ or $e \notin s$. Now, let C denote the collection of all conceivable states of a system. It follows that $\forall s \in C,\ s \subset \mathbf{\Omega}$. Note, we naïvely permit ourselves to use both logical operations *and* set operations on s, C, $\mathbf{\Omega}$, etc., though we do not claim that these collections are well defined sets, or sentences of a logic. Still, we will use logical operators on events so that every syntactically well-defined expression consisting of events and logical operators is itself an event (that may obtain in line with whether the expression is logically satisfied). E.g. since any state s is simply the conjunction of all the events that obtain, a state itself can be viewed as an event. Also, an event e occurs in a state s, iff s logically entails e: $e \in s \Leftrightarrow (s \Longrightarrow e)$.

Now, it is important to explain what we mean by *conceivability*. A state s is said to be conceivable (or a collection of events c are said to be jointly conceivable) if *every* event $e \in s$ which obtains in s (or in c) is itself conceivable and is logically consistent with every other event that obtains in s (or in c). Two events (or collections of events) are logically consistent if either event obtaining does not semantically/logically entail the negation of the other event. A single event is inconceivable if it is self-contradictory (e.g. the event that there are 2 passengers *and* 3 passengers in a car) or if it concerns objects which, we assume cannot exist (e.g. a neutron the size of a baseball). In a similar vein to our discussion on logical/sematic distinctness in the context of counterfactual causation, we posit that logical consistency is different from causal consistency. For example, having a car drive on flat terrain (with no strong wind or nearby electromagnets, etc.) may be causally inconsistent with the car having no fuel, but these events are not logically inconsistent, as we can invoke a miraculous force (à la David Lewis) that could push the car around. On the other hand, the event of the car having 30 l of petrol in its fuel tank is logically/semantically inconsistent with the event of the car being a fully electric vehicle; so, a state where both these events obtain is inconceivable. Alternatively, we may say that having a fully electric car with a half-filled petrol fuel tank is an inconceivable event, and so any state with this event is, by extension, an inconceivable state. The key take-away should be that the conceivability of some state s, in the sense that we have defined it, is not contingent on the causal or mechanistic *plausibility* of the collection of events that obtain under s, but rather contingent on whether the collection of events that obtain are jointly conceivable, i.e. logically/semantically consistent/reconcilable. Some further examples may help clarify our setting. E.g. Specifying a car's state entails specifying, among many other things, the amount of fuel in the car's fuel tank, which can be done at the granularity of litres or, in theory, at the level of individual hydrocarbon molecules. Now, due to the looseness of our informal definition of $\mathbf{\Omega}$, its elements may be logically compatible or mutually exclusive, e.g. let Ω_1 be the event that the car has a flat tyre in a front wheel and let Ω_2 specifically indicate that there is a flat tyre by the driver's side. Clearly, $\Omega_2 \Rightarrow \Omega_1$ even though these events are not equivalent. Moreover, Ω_1 is compatible with the event Ω_3 defined as the presence of a flat tyre in a back wheel, and so Ω_4, defined as the conjunction of Ω_1 and Ω_3, is conceivable. Not every conjunction of conceivable events is itself necessarily a conceivable event. E.g. if someone is holding up 3 fingers on their left hand (Ω_5); it cannot

concurrently hold true that they are holding up no fingers (Ω_6) on their left hand. Therefore, a hypothetical state s where Ω_5 and Ω_6 would both obtain is an inconceivable state and the event Ω_7, taken as the conjunction of Ω_5 and Ω_6, is an inconceivable event. As an aside, note that disjunctions of events are conceivable if at least one event in the disjunction is conceivable; however, we will be mainly focusing on conjunctions of events, and less so on disjunctions. Finally, the negation of an inconceivable event is trivially conceivable while the negation of a conceivable statement may still be conceivable. Indeed, our notion of conceivability is similar to the idea of satisfiability in logic and inconceivability is close to the notion of a contradiction.

Now, we wish to engage in the study of causal structure, and so we need to do better than just talk about the conceivability of events and states. Granted a system with some causal mechanisms and a conceivable event $e \in \mathbf{\Omega}$, we may consider the subcollection $S_e \subset C$ of all conceivable states where e obtains, and where the causal mechanisms of the system are *otherwise* respected, i.e. where all other events besides e and the events that are logically entailed by e, are *plausible* (i.e. do not contravene against any causal mechanisms that govern the system) granted a priori, that e obtains. Note that once we assume a state $s \in S_e$, for any event $x \in e$, no causal influence from other event in s may negate x. Also note, by stating that "$S_e \subset C$", we assume that the causal mechanisms of a system never generate inconceivable events. In other words, all plausible events are conceivable. Yet the converse may fail, e.g. events logically entailed by some conceivable e may be individually or jointly implausible (e.g. let e entail that a car is flying in a vacuum, or that there is no driver or autonomous driving system $[e_1]$ but that the car is steering itself $[e_2]$). We may call S_e the collection of all *plausible states* with respect to a given event e. In practice, it will not be possible to specify what states are causally consistent/plausible given some event e if we do not already know the causal mechanics of the system in detail (i.e. if we don't already know something we may wish to learn about); however assuming there exist some *latent* collection of causal mechanisms, i.e. a collection of well-defined rules for the system that govern what events cause what other events, we can posit that S_e, however intractable, is well-defined. Furthermore, let S without a subscript be the general unconditional collection of all plausible states (i.e., conceivable states that respect the system's causal mechanics) at instance t. Returning to the example of a car: consider a state s in which it is driving (event e_1) but suppose that it is doing so without fuel (call having no fuel "e_2"). A state s where $e_1, e_2 \in s$ is, as stated conceivable, but barring any miraculous scenarios (and perhaps, some pertinent external conditions, such as there being no steep slope or extreme wind, etc.) not plausible, i.e. $s \in C$ but $s \notin S$. Alternatively, let S_{e_1} be the collection of plausible states for which, e_1 obtains as a prior, and where the causal mechanics of the system are otherwise respected; then it follows that $\forall s \in S_{e_1}, e_2 \notin s$.

1.3.2 The measurement of systems

The vastness, vagueness and redundancies of $\mathbf{\Omega}$ preclude any construct directly derived from $\mathbf{\Omega}$ from being a practical setting for the study of causal structure. Even the setting of S where for any state, all events are logically and causally consistent, is too informal and abstract for any practical use. So, instead of keeping with the nebulous concept of a system state s (at some instance t), we will, as is done in practice, evaluate the state of any system via a collection of hopefully well-chosen *variables*, each of which, for any given s, yields a *measurement* of s.

In line with standard terminology, a system's state can then be approximately parametrised by a finite set $\boldsymbol{V}$ of n many *endogenous* variables (Hitchcock, 2023). Explicitly, we define an endogenous variable V_i, where $i \in \{1, \dots, n\}$, as the output of a mapping rule R_i that yields for *any* state $s \in C$ a real number v_i (the measurement) as the value for V_i, that is $V_i := R_i(s)$. In particular, for any value $v_i \in \mathbb{R}$ that V_i may assume, there will exist a collection $c_{v_i} \subset \boldsymbol{\Omega}$, such that $\forall s \in C, e_{v_i} \in s \wedge e_{v_i} \in c_{v_i} \Longrightarrow R_i(s) = v_i$. Indeed, if $e \in c_{v_i}$ for some $v_i \in \mathrm{Range}(R_i)$, we say that V_i measures the event e. Importantly, for any distinct pair of measurements a and b for V_i, events sufficient for obtaining either value must be logically incompatible:

$$(a \neq b) \wedge (e_a \in s \Longrightarrow R_i(s) = a) \wedge (e_b \in s \Longrightarrow R_i(s) = b) \Longrightarrow e_a \wedge e_b \notin \boldsymbol{\Omega}$$

We stress that for an endogenous variable V and for its corresponding mapping rule R to be well-defined, *every* conceivable state must yield some measurement, that is $\forall s \in S, \exists\, r_s(s) \in \mathbb{R}\; s.t.\; R(s) = r_s$. So, every endogenous variable partitions C into disjoint subcollections. As a concrete example of an endogenous variable, let V_1 correspond to the rounded value indicated on a thermometer in a car. The value of V_1 constrains s by ruling out states that are incompatible with a given measurement, and so we gain information about the actual state s if we learn the value of V_1. If we observe $V_1 := 20$, then s must be a state where a thermometer reading near the value $20°\mathrm{C}$ likely obtains, as via *additional* causal reasoning, we may *assume* that $c \notin s$, where c is the event that the actual temperature inside of the car is 40°C. In contrast, we may not rule out the event r that the temperature inside of the car is near 20°C, as this seems to be causally consistent with the measurement. Also, as stated, every variable $V_i \in \boldsymbol{V}$ inherits a form of *internal logical consistency* from s, via the requirement on the pairwise mutual exclusivity of events that induce different measurements, e.g. let V_2 count the total number of passengers in the car and note that it can only assume a single value for any given s. Do note that such internal logical consistency is only assured if our variables are well defined, as we can conceive of ill-defined variables: let V_i be defined such that $V_i := 1$ is the required assignment if our car has automatic transmission, and $V_i := 2$ if our car has no clutch pedal. Granted a typical automatic car, the value of V_i then cannot be determined as it must simultaneously evaluate to both 1 and to 2. Essentially, internal logical consistency equates to the requirement that R_i is a well-defined function on C. On a different note, a notion of instance-indexing / time-indexing for s is inherited by $\boldsymbol{V}$ (as measurements take place with respect to some time or instance), though as we will see in section 1.4.2, the interpretation of time vis-à-vis causality can be a delicate affair. Yet, while glossing over intricacies related to time, note that if an event $e_1 \in s$ is sufficient for some V_i to evaluate to a value v_1, then e_1 cannot possibly *cause* some event e_2 that would be sufficient for V_i to evaluate to another distinct value v_2. We say that all variables must also be *internally causally consistent*.

Moving on, recall from section 1.2.2 the criterion of distinctness formulated by Lewis for event pairs in a causal relation. We extend this criterion to variables by requiring that the variables in $\boldsymbol{V}$ are chosen in a manner so that for no subsets $\boldsymbol{A}, \boldsymbol{C} \subset \boldsymbol{V}$ and sets of values a, c, it follows that $\boldsymbol{A} := a \Longrightarrow \boldsymbol{C} := c$, unless $\boldsymbol{A} \equiv \boldsymbol{C}$, and, by extension $a \equiv c$. That is, there are no valuations for any variables that logically/semantically entail valuations for some other variables. This condition can be violated: e.g. "$V_3 :=$ number of passengers in the front row of the car" and "$V_4 :=$ number of passengers in the back row" may stand in a logical relation with the variable V_2 counting the total number of passengers. Indeed, any entailed valuations between variables in $\boldsymbol{V}$ should follow only by virtue of some latent causal dependencies/mechanisms.

1.3.3 Atomic and amalgam causal mechanisms

We are now at an appropriate stage where we may readdress the notion of causal mechanisms. Let $c \subset \mathbf{\Omega}$ be a collection of jointly conceivable events. Also, let w be a collection of jointly conceivable events that are *external*, i.e. beyond the scope of the system (i.e. $w \not\subset \mathbf{\Omega}$). Now, let $e \in \mathbf{\Omega}$ be an event in the system that is not semantically/logically entailed by c, i.e. $c \not\Rightarrow e$ (by construction, events such as w that obtain outside of the bounds of the system cannot logically entail an event e inside of the system, because the "outside" and "inside" are separated regions of spacetime). There exists an *atomic mechanism* from c and w to e, if there exists an appropriate event $p \in \mathbf{\Omega}$ called a *context* such that:

(B) For any $s \in S_p$, if c and w obtain, then e obtains and *every* event in the collections c and w is a cause of e in the counterfactual sense.

Now, there is only one requirement on p for it to serve as valid context: namely that the conjunctions $c \wedge e \wedge p$, $\neg c \wedge e \wedge p$, $c \wedge \neg e \wedge p$ and $\neg c \wedge \neg e \wedge p$ must all be conceivable and constitute *fully* specified system states. This essentially means that p does not logically entail anything about c and/or e (and vice versa), so it is strictly a specification of all events that c and e do not logically/semantically entail or negate. Furthermore, it means that p must serve as a specification of *all* the aspects of the system that are not specified by c and e.

Note, granted some c, w and e, there may be more than a single valid context with respect to which definition (B) is satisfied. Therefore, our perspective on mechanistic causation is similar to the influential INUS (*insufficient but necessary part of an unnecessary but sufficient*) condition for causation given by J. L. Mackie (1980), with a distinguishing characteristic being our requirement on an explicit reference to a context. Now, granted a context p for some atomic mechanism, consider every event $r \in p$ s.t. $\forall q \in \mathbf{\Omega}$ where $(p \setminus r) \wedge q \Longrightarrow \neg r$ and where $(p \setminus r) \wedge q \wedge c \wedge e \in C$, it follows that "$(p \setminus r) \wedge q$" is a context for the same atomic mechanism. In turn, let $k \subset p$ be the "part" of p that is independent of every such r (i.e. where $\forall r$, $k \wedge r$ and $k \wedge (\neg r)$ are both conceivable) and where k is *maximal* (for every other $k' \subset p$ that is independent of every r, it follows that $k' \subseteq k$); we call k the *core* of the context p. It lies beyond the scope of this thesis, but if $k \in s$, it follows that c and w are *necessary* and *sufficient* causes for e in a counterfactual sense. Therefore, k is *the* salient aspect of the context p that effectuates the atomic mechanism. Accordingly, knowledge of any event $r \in p$ that is not entailed by k is redundant for determining the atomic mechanism and the realisation of e. Note, k may be the empty set if c and w are, a priori, necessary and sufficient causes for e. Now, a supplementary concept that we will define is called the *collection of cores* (denoted by σ) for a given atomic mechanism from c and w to e. Suppose that c and w, under a core k, specify an atomic mechanism for an event e. However, for the same c, w, and e, there may be other distinct cores k', k'', ... derived from distinct contexts p', p'', ... under which definition (B) is satisfied. The collection of all such cores is σ, and while σ itself is not a core, every element in it is. Note, if the empty set is a core, then it follows that σ is singular.

Moving on, note that under the hood, we assume a notion of *regularity* or physical *determinism* in order to define atomic causal mechanisms; this follows from our claim that obtaining $c \wedge w$ will *always* yield, through the atomic causal mechanism, the event e (granted context p). We therefore rule out from consideration any fundamentally non-deterministic mechanisms, such as those that may be encountered in quantum mechanics (Woodward, 1989). However, we

consider the limitation to deterministic settings as a useful simplification granted the abundance of systems that can be regarded as deterministic (e.g. macroscopic systems).

Since conjunctions of logically consistent events are events, we may regard c and w as a single event d, so there exists an atomic mechanism from d to e. Let c_e be the collection of all events such that $\forall d \in c_e$, there exists an atomic mechanism from d to e (with a corresponding collection of cores "σ_d" for each d). We may denote the collection of all mechanisms from c_e to e as the *amalgam mechanism* or simply as *the* mechanism for e. Furthermore, we may relax the view of e as a single event, and instead allow e to be a collection of events of interests, and we may consider the collection of all amalgam mechanisms that lead to *some* element of e as a single amalgam mechanism. Therefore, we have 2 types of mechanisms: atomic ones that directly satisfy our definition (B), and amalgam mechanisms that are simply collections of atomic mechanisms. We consider if there are issues with generalising our view of e as being a collection of events: what if 2 (or more) distinct events e_1 and e_2 within e share a common cause. That is, granted some c, consider the case where there exists a mechanism from c to e_1 and a distinct mechanism from c to e_2, e.g. the event c of smoking in a room may cause a smoke detector to go off (the event e_1) and another person in the room to cough (e_2). This situation seems reasonable. Indeed, our only expectation is that different events that may jointly obtain due to distinct mechanisms and a common cause cannot be logically inconsistent. E.g. the event q of putting a thermometer in a cup cannot cause the temperature reading to both rise (the event e_1) and fall (incompatible with e_1) simultaneously. In review, given a collection e of the form $\{e_1, \dots, e_n\}$, with n many events of interest, we define the amalgam mechanism for e as follows: For every event e_i in e, identify all atomic mechanisms leading to it, and for each event, note its causes from within the system and beyond, as well as the collections of cores. Now, the main category of amalgam mechanisms that we will consider are those for collections of mutually exclusive events: i.e. let e be a collection of the form $\{e_1, \dots, e_d\}$ of d many events of interest. The collection consists of mutually exclusive events if $\forall e_i \in e,\ e_i \Longrightarrow \forall e_j \in e \backslash e_i\ (\neg\ e_j)$, i.e. the only conceivable conjunctions of the events are of the form: $(\neg e_1) \wedge \dots \wedge (\neg e_{i-1}) \wedge e_i \wedge (\neg e_{i+1}) \wedge \dots \wedge (\neg e_d)$.

We now review our definitions for atomic mechanisms an example. Let some hypothetical system K be fully specified by 3 events A, B and C that are logically compatible in any *configuration*. That is, there are no logical/semantical statements such as $A \wedge B \Longrightarrow \neg C$ that are valid. Then, by our definition of conceivable states, there are 8 conceivable system states. Also, assume the existence of these atomic mechanisms:

1. A must obtain if an event W_1, which lies beyond the scope of the system, obtains.
2. A must obtain if an event W_2, also beyond the scope of the system, obtains.
3. B obtains if A obtains.
4. C obtains whenever both B and an event W_3 beyond the scope of the system obtain.

We will furthermore require that the external events W_1, W_2, W_3 are also jointly conceivable in every configuration. This ensures that the full repertoire of mechanisms in K can be elicited. Figure 1 presents K, with arrows pointing from events that feature as causes in some atomic mechanism towards events that feature as the corresponding effects:

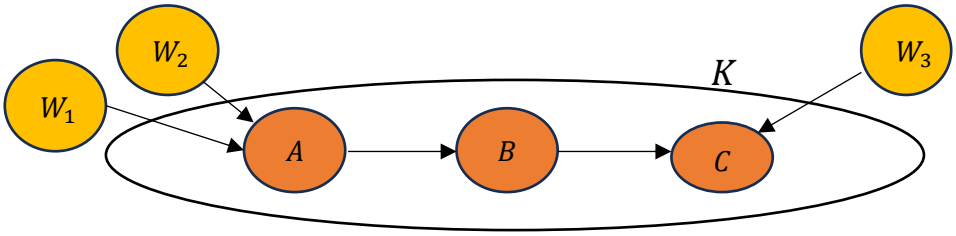


Figure 1. The system "K"

Based on our definition (B) of atomic mechanisms, we may imagine that there exists an atomic mechanism leading from A and W_3 to C. After all, it seems that C obtains in every *plausible* state whenever A obtains (granted W_3 obtains), because A causes B, and B and W_3 jointly cause C. I.e. granted that $w = \{W_3\}$ obtains, the occurrence of C counterfactually depends on the occurrence of A. However, while this pattern of dependence is guaranteed to hold for any plausible state (granted the mechanisms that we have described), we have not specified a context p, and this seemingly small technicality is what voids the spurious atomic mechanism, as we will see that it is impossible to find a valid context. Consider S_B, the collection of all plausible states granted the that B obtains. Note that "B" amounts to a possible context. Now, granted B, A obtains depending on whether at least one of W_1 or W_2 obtain; also, regardless of whether A obtains or not, the only event that will determine whether C obtains is W_3 (since B already obtains by assumption). Therefore, under S_B, we have a plausible state where A, B and C obtain, but it is also clear to see that C does not counterfactually depend on A (since B, by assumption obtains regardless of A). In other words, we have found a possible context $p = B$, as $A \wedge B \wedge C$, $\neg A \wedge B \wedge C$, $A \wedge B \wedge \neg C$, and $\neg A \wedge B \wedge \neg C$ all constitute conceivable and fully specified states, but our condition (B) for an atomic mechanism leading from A (and W_3) to C is not satisfied under p. However, we are not finished, as we must check that definition (B) fails under *every* possible context. Thankfully, there is only 1 other possible context (namely "$\{\neg B\}$"), and it is easy to check that C cannot obtain independently of A, disproving the existence of the atomic mechanism under question. In general, we see that proving that an atomic mechanism does not exist can be harder than showing that one does exist, as the former requires checking *all* possible contexts. Also, note there exists an atomic mechanism from A, W_3 *and* B to C. Let our context p be empty; since A, B and C fully specify K, this is valid. Clearly C counterfactually depends on B and W_3. Crucially, C also counterfactually depends on A, since negating A causes B not to obtain, which causes C not to obtain.

1.4 Deterministic Structural Causal Models (SCMs)

1.4.1 Exogenous variables and SCMs

Having spent some effort to define causal mechanisms, we will now fruitfully marry our ideas on amalgam causal mechanisms with the idea of a set of endogenous variables $\boldsymbol{V}$ that measure a system's given state s. Let's begin by noting that in practice, we almost certainly will not have the right set of variables that will allow us to exactly represent, let alone identify any *true*

underlying mechanisms that a system obeys. Possibly except for a few high-level discrete properties of a system, e.g. V_2 (the number of passengers in a car), we will only have coarsened and approximate *proxies* that give us limited information about the true *latent state* of the system. E.g. a high reading on a thermometer is in no sense the *true* cause of passengers sweating in the car, rather the thermometer measures some event that obtains in state s, that is presumably, at least in part, involved in the mechanism driving the sweating. I.e. the variable V, indicating the thermometer's reading, is a good proxy for air temperature, and if we are to utilise V in some causal formula for determining how much the passengers sweat, we must remember not the take the formula too literally, but rather to view each variable in the formula as a seemingly useful proxy for some aspect of the true latent state of the system.

Taking matters further, there may also be many completely unmeasured events that play a decisive role in whether a measured event of interest occurs (e.g. that are involved in atomic causal mechanisms that cause an event sufficient for mapping V_i to some value v_i), and these unmeasured events may wholly, or in part, be external to the system, e.g. a car tyre getting punctured, depends not only on the tyre's condition, but also on the presence of sharp objects on the road that the car is on; The sweating of passengers depends, in addition to the air temperature, on the air humidity in the car, the passengers' stress levels, the amount of clothing they are wearing, etc., all of which we may not be measuring via any endogenous variables. Or perhaps, we may be measuring pertinent events, but not at a sufficient *resolution* that the applicability of a causal mechanism can be determined (e.g. if V_i tracks temperature but a resolution of 10° blocks, the available information may be too coarse for us to determine whether passengers are sweating or not). So, to factor in limitations to our knowledge with regards to possibly pertinent events within the system and beyond it, which we have not measured, but which may nonetheless feature in some causal mechanisms of interest, we invoke the notion of *exogenous variables*, which, *by construction*, are assumed to measure and discriminate between *many* such pertinent events up to a resolution that makes the applicability of *many* pertinent mechanisms determinable. Note that exogenous variables seem related to external events $w \notin \boldsymbol{\Omega}$, which we introduced to specify those events that are a part of some atomic mechanism, but which lie beyond the scope of $\boldsymbol{\Omega}$. However, a key difference is that exogenous variables are not restricted to measuring external events and may measure any event, external or not, that simply lies beyond the scope of what can be measured via $\boldsymbol{V}$.

Now, for every variable $V_i \in \boldsymbol{V}$, we may be interested in the possibly numerous atomic mechanisms that govern the realisation of the events that are measured by V_i. Strictly speaking, distinct atomic mechanisms may cause events that yield the same value for V_i, so for the sake of simplicity, we allow ourselves to deal with the amalgam mechanism that governs the occurrence of every event that corresponds to every possible value of V_i. For example, a reading of 25°C for the temperature of the car may be obtained due to body heat from passengers sitting the car, the car's AC system warming the air, 25°C simply being the ambient temperature around the car, any combination of these factors, or because the car's thermometer is somehow broken and stuck at that value, etc., and we may consider as the amalgam mechanism, all possible atomic mechanisms that, with under some core from a collection of cores, could yield a particular temperature measurement, for every value of the thermometer.

Every event that corresponds to a measurement / a value for V_i, is itself, presumably causally contingent on various possible collections of causative events and cores, many constituents of which we will have hopefully measured via $\boldsymbol{V}\backslash V_i$, if indeed the set of variables in $\boldsymbol{V}$ have been

well chosen. Now, we can define a function f_i that numerically *models* the amalgam mechanism for the collection of all events that correspond with values for V_i, based on what values have obtained for the rest of $\boldsymbol{V}\backslash V_i$, while also factoring in unmeasured pertinent events (e.g. if a certain cause, core or outside event obtains) via an exogenous variable U_i. Again, we stress the importance of ensuring that our endogenous variables in $\boldsymbol{V}$ are well chosen, meaning that the amalgam mechanisms leading to the events measured by V_i *generally* involve causes that are events measured by $\boldsymbol{V}\backslash V_i$. If this is not the case, there will be little causal structure to analyse among the variables of $\boldsymbol{V}$, making the study of the system via the variables of $\boldsymbol{V}$ essentially fruitless. Now, repeating the construction of modelling functions for every i yields the formalism of structural causal models (SCMs), also known functional causal models (FCMs). In line with Pearl (2009, p. 27), an SCM then is defined as the triple $\langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{f} \rangle$, where $\boldsymbol{f}$ is a set of mappings $\{f_1, f_2, \dots, f_n\}$ of the form $\boldsymbol{f}: \boldsymbol{V} \times \boldsymbol{U} \to \boldsymbol{V}$ that models the corresponding amalgam mechanisms for the endogenous variables in $\boldsymbol{V}$, i.e.:

$$V_i := f_i(PA_i, U_i), \quad i \in \{1, \dots, n\}$$

Where $\forall i \;\; PA_i \subset \boldsymbol{V} \setminus V_i$, and where we have as stated, a single $U_i \in \boldsymbol{U}$ for each mapping f_i that measures, to a sufficient resolution *many* pertinent events that have either not been measured at all, and/or simply not sufficiently well for determining V_i. Under specific circumstances, e.g. when an SCM is invoked in the philosophical literature as a tool for the analysis of causality, $\boldsymbol{U}$ may be observable in the sense that the values of every variable $U_i \in \boldsymbol{U}$ are predetermined and known to us. So, under such a scenario where every exogenous variable has an observable value instead of being an unknown, the value of every variable in $\boldsymbol{V}$ should be computable and hence "determinable", hence an SCM where $\boldsymbol{U}$ is observable is called a *deterministic* SCM (Hitchcock, 2018). In practice, the observability of $\boldsymbol{U}$ can be a fanciful assumption, as it demand a great deal of certain knowledge about many events in the system and beyond it. Furthermore, the observability of $\boldsymbol{U}$ arguably defeats the idea of defining it as *the* pertinent events/information that cannot be measured. Therefore, deterministic SCMs will only serve us in setting the stage for non-deterministic SCMs, which are more tenable. As an aside, we gave the example of the philosophical literature as a domain where deterministic SCMs are studied, however this was somewhat disingenuous. What's more, we would like to acknowledge that the association we have promoted between exogeneity and unobservability does not hold in all modelling frameworks; e.g. in structural equation models (SEMs), and in many disciplines across the social sciences, psychology and economics, an exogenous variable need not be unobserved or unmeasured, rather exogeneity is often meant with reference to the model itself, i.e. a variable is exogenous (as opposed to endogenous) if there is no function within the model for estimating it or solving for it, but we may still have access to its value (Engle, Hendry and Richard, 1983; Woodward, 1995; MacCallum and Austin, 2000).

In line with the literature, we will refer to "$V_i := f_i(PA_i, U_i)$" as the *structural equation* (Menzies and Beebee, 2020) for determining the value of the endogenous variable $V_i \in \boldsymbol{V}$. Strictly speaking, the symbol ":=" is not an equality, but rather an assignment operator. The usage of assignments coheres with the asymmetry of causation: while causes determine their effects, there is no corresponding default backwards causation. E.g. the rain making the lawn wet does not imply that watering the lawn will make it rain. Furthermore, every $f_i \in \boldsymbol{f}$ not only indicates a direction of causation, but also identifies the exact set of other endogenous variables PA_i called the *parents* of V_i, which alongside U_i, measure events that serve as the

causes or cores for the events measured by V_i. Indeed, assuming that some f_i, is accurate, it follows that if some variable $V_j \in \boldsymbol{V}$ is not an argument of f_i, i.e. if $V_j \notin PA_i$, then no events measured by V_j can appear as a cause in *any* atomic mechanism that causes an event that is measured by V_i. We forego a proof, but this result ensures that for every i, f_i represents *all* the mechanisms relevant to V_i. Again, it is worth stressing that an accurate $f_i \in \boldsymbol{f}$, aside from being an effective computational tool for predicting measurements, may be regarded as a summary of what variables $PA_i \subset \boldsymbol{V} \setminus V_i$ actually measure events that feature as causes or even as cores for an atomic causal mechanism that determines an event measured by V_i.

In practice, SCMs are often studied in conjunction with interventions, which can be regarded as manipulations of an underlying system that affect the workings of some mechanisms. A *perfect* or *surgical intervention* modifies the system in a manner that completely overrides the actual mechanisms (detailed by $\boldsymbol{f}$) for precisely some subset $\boldsymbol{T} \subset \boldsymbol{V}$, such that the variables in $\boldsymbol{T}$ instead assume some targeted values, while the amalgam mechanisms and corresponding structural equations for other variables in $\boldsymbol{V} \setminus \boldsymbol{T}$ are preserved. Perfect interventions are theoretically supported by an assumption that systems are governed by *autonomous* (Pearl, 2009, p. 28) or *independent* mechanisms that "do not inform or influence one another" (Parascandolo et al., 2018), so intervening on some mechanisms should not have to inevitably disrupt other ones (Goyal et al., 2020). Henceforth, interventions are assumed to be perfect unless stated otherwise. Now, recall the idea of a context p from definition (B), which by itself, can be an implausible event. The framework of interventions allows us to theoretically justify the assumption of implausible contexts, with the idea being that a context can be imposed via an intervention, and so we only require that our ability to intervene in a system is substantial enough that we can impose any conceivable context. There is, however, an argument against this line of thinking: presumably, interventions are not realised via miracles but by exploiting some causal mechanisms of nature that can influence the system; therefore we should not be able to realise any implausible contexts, or rather, it follows that every conceivable state is in fact plausible (by virtue of some appropriate intervention) and so, by extension, there is no such thing, per say, as an implausible state. This may present a problem, because our definition (B) of an atomic causal mechanism equates to the idea that if some cause c obtains (and possibly the right core σ and external conditions w), then some effect e obtaining is the *only* plausible option, yet what if there were an additional concurrent intervention that blocks e from obtaining? This seems plausible since we need not contravene against any physical law, and yet we can seemingly void definition (B) for any atomic causal mechanism (as we can negate an anticipated effect via an intervention). A way out of this apparent quagmire lies in first accepting that if it is true that we can intervene in such powerful ways in some given system, then indeed every conceivable system state does in fact become plausible. However, the existence of an interventional mechanism for controlling some event or variable, for the matter, does not negate that there could be other mechanisms that could cause or negate the event. Rather, it forces us to accept that certain mechanisms (for example, interventional mechanisms) can *override* other mechanisms. In other words, there must be a *hierarchy* of mechanisms. E.g. eating ice cream may cool our body down while being in the sun warms it up; yet eating ice cream outdoors on a sunny day does not lead to a contradiction; rather, the effects of one mechanism simply overwhelm the other. Therefore, if we wish to show the existence of an atomic mechanism, we must ensure, via σ and by possibly extending what events are in w, that we negate any events that could amount to an intervention, or in general,

to any mechanism that could override the mechanism of interest.

Another question that may come up is what exactly the exogenous variables $\{U_1, , \dots, U_n\}$ are meant to account for from all the possible events that could play a role in determining the values that are obtained for $\{V_1, , \dots, V_n\}$, but that are not measured at all, or at least well enough by $\{PA_1, \dots, PA_n\}$. In our definition of $\boldsymbol{U}$, we demanded that $\boldsymbol{U}$ measures *many*, but not *all* the possibly pertinent unmeasured events, and this was done for good reason. If $\boldsymbol{U}$ is meant to model *every* such event, then, it must in advance also model every event corresponding to some hypothetical intervention that could be realised. Yet, interventions were defined as contraventions against the mechanisms modelled by $\boldsymbol{f}$ (and by extension by $\boldsymbol{U}$). This, by itself, appears to lead to a contradiction. Now, if we try to rectify $\boldsymbol{f}$ and $\boldsymbol{U}$ in order to account for all possible interventions, we would end up, for any i, with a function f_i that is never constrained under any realisation of the parent variables of V_i, because by assumption, a perfect intervention could still impose *any* targeted value for V_i under any values for the parents PA_i. This issue does not hold if $\boldsymbol{U}$ only measures "*many*" of the possibly pertinent events that are not measured by $\{PA_1, \dots, PA_n\}$. In particular, we do not expect that $\boldsymbol{U}$ "pre-measures" any purposeful interventions that we ourselves carry out on the system. Nonetheless, what is measured by $\boldsymbol{U}$, must, in some sense, be broad enough that for an un-intervened system, we can ignore what $\boldsymbol{U}$ does not measure, as only then can we hope to predict every $V_i \in \boldsymbol{V}$ via f_i.

Moving on, it follows that for an SCM to be a generally accurate representation of an underlying system, it must be true that the constraints implied by the SCM's structural equations are a *morphism* of the structure of some of the real causal mechanisms of the system. Explicitly, given a system state that implies certain values for some subset PA_i of $\boldsymbol{V}$, and granted some value for U_i, it should follow that an event holds that precisely maps the endogenous variable V_i to the value "$f_i(PA_i, U_i)$". Furthermore, by assumption, every V_i is, via f_i, determined by its *direct causes* or *parent* variables, that is, by PA_i together with the exogenous variable U_i, and so it follows that every other variable $X \notin \{PA_i \cup U_i\}$ has no bearing on V_i w.r.t constant values for the parent variables of V_i. To elaborate: knowing the value of X may help us infer V_i, especially if we have incomplete information on the values of PA_i and U_i, or if it turns out that "$f_i(PA_i, U_i)$", though referencing the right variables, is a numerically inaccurate formula, yet, as the value of V_i is not directly determined by X, a change to X, e.g. via an intervention, cannot affect V_i unless the values of the parents of V_i are changed. One may wish to criticize this as being tautological, as it is only natural that a perfect intervention that exclusively manipulates the value of X and does not interfere with any other mechanisms that determine the values of other variables, should, by definition, not affect other variables such as V_i. However, such a criticism indicates a misunderstanding of how SCMs work: an intervention that targets X, can, in line with the causal dependencies implied by the structural equations, change the values of X's *child variables* (i.e. variables that have X as a parent, as their values are determined via the structural equations by the value of X), or indeed influence any $De(x)$, that is, any *descendent variables* of X (the children of X, the children of the children of X, etc.), and it may very well be true that $V_i \in De(X)$, (i.e. $X \in An(V_i)$, X could be an ancestor of V_i), in which case a change in X may result in a change in V_i, though this would have to be *mediated* via a change in the values of some variable in PA_i.

1.4.2 Time and acyclic SCMs

We are at an appropriate point to address an elephant in the room: namely, how to interpret time in the context of SCMs. Consider a simple system consisting of a freezing room with a heating radiator and an ice cube that is placed near the radiator. Presumably, there will be causal mechanisms leading from the radiator to the ice cube that will cause the melting of the ice cube (e.g. heating via direct IR radiation, via convective currents of warm air, etc.). However, if we take a snapshot view of the system, perhaps right after turning on the radiator, it will *appear* that the radiator has no causal influence on the ice cube. Of course, we do not interpret this snapshot as proof of there being no causal mechanisms. Rather, we know that causal effects need time to *propagate*, and that in many instances, for there to be any significant or measurable effect, this time can be in order of seconds, hours or even years or longer (e.g. the symptoms of a cancer may only after decades of continual exposure to a carcinogen). The nuances of time are even greater for systems featuring cyclical patterns of causal dependence. On the one hand, it is a standard assumption that causation should respect time order, in the sense that causes should precede their effects and that the present (or future) cannot change the past (Frisch, 2022). On the other hand, we can imagine 2 variables that can influence each other, e.g. populations of predator and prey, the supply and demand of goods, or in general, processes featuring feedback, which e.g. appear abundantly in biology (Rehder, 2017). Indeed, the formalism of deterministic SCMs that we have presented only allows us to compute the value of any given V_i, granted that we know the values of U_i and V_j $\forall V_j \in PA_i$, and to its credit, this principle coheres with the notion that causes first have to obtain before their effects may obtain or become determinable. Yet, what if for some i and j, V_i and V_j both measure events that feature in each other's respective mechanisms, and thereby also feature in the set of each other's parent variables: there appears to be no obvious way to solve this chicken-and-egg problem. A seeming approach could be to treat the structural equations as a system of simultaneous equations that are to be solved. However, it is unclear how to proceed if no unique solution or an infinite set of solutions are found. On the other hand, some systems, under some value assignments for $\boldsymbol{U}$, may have no numerical solutions whatsoever. How are all these scenarios to be interpreted? Furthermore, we developed SCMs as a part of a framework for causality; it is not clear if simply deriving numerical solutions to the system of equations derived from an SCM presents a conceptually valid procedure for determining what actual measurements from the underlying system should obtain.

So let us briefly discuss how issues stemming from cyclicity are approached in the literature. Firstly, we may note that if one the variables were to be determined instead by an intervention, then the chicken-and-egg issue would be resolved, and we could use the remaining unperturbed function to compute the remaining undetermined value (e.g. use f_i to determine V_i now that we know V_j). Indeed, this approach preserves the interpretation of the causal structure suggested by the functions in $\boldsymbol{f}$, and its theoretical implications have been exploited, e.g. to study the statistical properties of certain simple linear systems with patterns of cyclical dependence (Spirtes, 1993) and, as of recent, also non-linear cyclical systems (Forré and Mooij, 2017). However, that's not all; recall the notion that causal influences need time to propagate: in line with this principle, whenever we consider a state s and the value of some V_i in s, note that the function f_i which computes V_i, should, strictly speaking, utilise as arguments the *past* values of other variables, as only events that lie in the past can have possibly

influenced a current event that obtains now (and has given us the current value of V_i). So, given some time index t, a more sensible formalism for SCMs would then look like:

$$V_i^{t+1} := f_i\left(PA_i^{t+1}, U_i^{t+1}\right), \;\; i \in \{1, \dots, n\}, t \in \mathbb{N}$$

Where $PA_i^{t+1} \subset \{V_1^t, V_2^t, \dots, V_n^t\}$. Note, that now we can even have "self-causation" where the value of a variable plays a role in determining its own next value. If we adopt this formalism, the possibility of causal cycles appears to be eliminated, and the causal meaning of each f_i is clearly interpretable. Indeed, many frameworks cohere with this formalism and have been extensively studied, with some influential examples being Dynamic Causal Models (DCMs) in neuroscience (Friston, Harrison and Penny, 2003; Friston, 2011; Friston et al., 2011), dynamic Bayesian Networks in machine learning (Ghahramani, 1998; Mihajlovic and Petkovic, 2001) and Granger causality in econometrics and statistics (Granger, 1969; Shojaie and Fox, 2022). Yet this formalism has, unsurprisingly, its own limitations. One possible practical limitation could be that that for real-world systems that we may wish to apply our framework to and study, there may simply not be any time-series data available. Also, real-world concerns aside, even in theory, when we define the size of the temporal gap corresponding to a time index increment, there may be the issue that some effects simply propagate faster than the chosen interval of time, leaving open the possibility for cyclical influences to still play out. In particular, if our system is a continuous dynamical system or best understood as such, then perhaps only by letting the temporal gaps approach length 0 can cyclical causal influence loops be curtailed. However, at that point, we are essentially engaging in the study of systems of ordinary differential equations. In short, whether due to practical or theoretical reasons, we simply may not have the ability to study systems at the temporal resolution required using our framework (Chicharro and Panzeri, 2014). On a different note, some effects may propagate very slowly and so require many time-increments before they become measurable. This too, can present an issue in practice, though at least in theory, it can be addressed by extending the span of what past variables may serve as causal parents, i.e. by letting $PA_i^{t+1} \subset \{V_1^t, \dots, V_n^t, V_1^{t-1}, \dots, V_n^{t-1}, \dots, V_1^1, \dots, V_n^1\}$.

As stated, some systems may be characterised as systems of ordinary differential equations or as stochastic differential equations. In such cases, SCMs with cycles may, instead of representing causal mechanisms, serve as compact representations of the solution spaces of the systems. That is, for every i, $V_i := f_i(PA_i, U_i)$ becomes an equation describing the value of V_i at equilibrium (Mooij, Janzing and Schölkopf, 2013; Bongers et al., 2021). However, in such frameworks, the causal interpretation of the functions in $\boldsymbol{f}$ as representations of the system's actual causal/dynamical mechanics, outside of the scope of simple linear dynamical systems (Fisher, 1970), can be lost. This is because the equilibrium equation and differential equation for any variable may involve different terms and have different forms.

Now, for our purposes we will take the simplest approach and assume, a priori, that we are studying systems without any cyclical patterns of causal dependence, and that we measure the system at a stage where all effects have had enough time to fully propagate. The structural equations then, give the final expected measurement for when the system has reached a *steady state*. So, from here on, SCMs are, by default, assumed to be acyclic. Importantly, the acyclicity condition implies a *unique* valuation for $\boldsymbol{V}$ w.r.t any $\boldsymbol{U}$, and allows us to draw a causal *directed acyclic graph* (DAG) to portray the SCM's causal dependencies. As a toy example, we present the "gas grill" SCM in Figure 2, adapted from Hitchcock (2023):

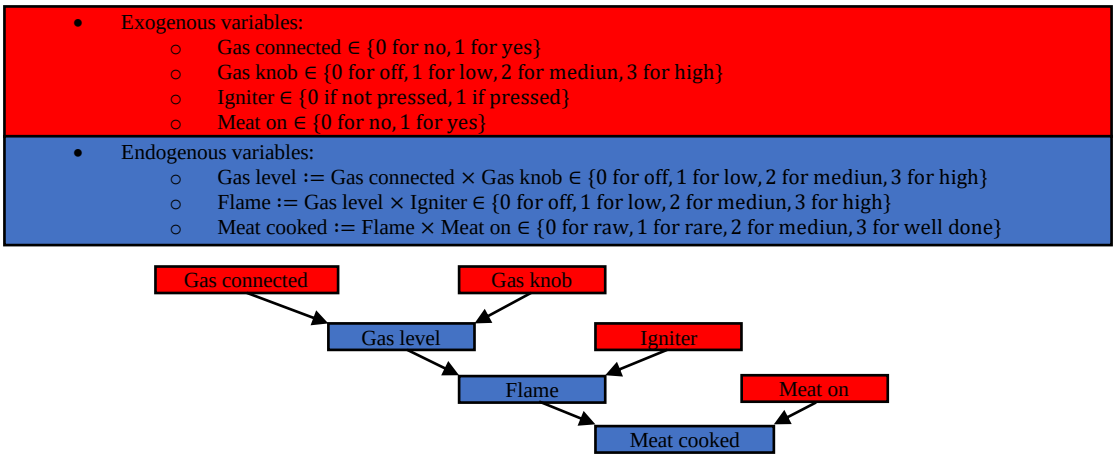


Figure 2. The acyclic "gas grill" SCM and its causal DAG

While the "gas grill" SCM specifies how endogenous variable in $\boldsymbol{V}$ is to be precisely computed, note that the causal DAG only specifies the parents of each variable. As such, SCMs are more informative than DAGs. Indeed, knowing an SCM allows us *simulate* states, by computing what $\boldsymbol{V}$ should be under any counterfactual settings, i.e. given any values for $\boldsymbol{U}$, or when some intervention is made that overrides the structural equations for a certain subset $\boldsymbol{T} \subset \boldsymbol{V}$. E.g. if we intervene to set Flame $:= 0$, e.g. by drenching the grill's burner in water, then it follows that Meat cooked $:= 0$. Also, we will have broken the chain of causation running from "Gas level" to "Meat cooked", so the value for "Meat cooked" will now be independent of e.g. whatever the value of "Gas knob" may be. It follows that acyclic SCMs, in conjunction with interventions, give rise to a logic of counterfactuals (Briggs, 2012) through which we may derive the expected effects of any values for $\boldsymbol{U}$ and any intervention. Conversely, we will know that an SCM is inaccurate if, under some $\boldsymbol{U}$ and/or interventions, it makes incorrect predictions. Consider the SCM $\boldsymbol{M}_1 = \langle\{V_1, V_2\}, \{U_1\}, \{f_1, f_2\}\rangle$, where $U_1 \in \{0,1\}$, $V_1 := f_1(U_1) = U_1$, and $V_2 := f_2(V_1) = 2 \times V_1$. Now suppose that the actual underlying causal mechanisms run instead from U_1 to V_2 and from V_2 to V_1, and cohere with the SCM $\boldsymbol{M}_2$ with functions $V_2 := f_2(U_1) = 2 \times U_1$ and $V_1 := f_1(V_2) = 0.5 \times V_2$. Note that without interventions, $\boldsymbol{M}_1$ correctly predicts the true system and, so is *observationally equivalent* to the correct SCM $\boldsymbol{M}_2$. However, $\boldsymbol{M}_1$ would incorrectly predict that an intervention that fixes $V_1 := 1$ would also set the value of V_2 ro 2. Yet V_2 does not, in fact, causally depend on V_1, and indeed if $U_1 = 0$ were to occur, we would observe $V_2 := 0$. Indeed, one can easily prove that *any* inaccuracy in an SCM can be exposed with the right intervention and set of values for $\boldsymbol{U}$.

In summary, deterministic SCMs may be viewed as concise predictive models, which take as input some observable $\boldsymbol{U}$ and an intervention on a subset $\boldsymbol{T} \subset \boldsymbol{V}$ ($\boldsymbol{T}$ may be empty) and via $\boldsymbol{f}$ predict values for $\boldsymbol{V}$. Again, only if $\boldsymbol{f}$ correctly represents the latent mechanisms, will the effects of *every* intervention under *every* possible $\boldsymbol{U}$ be accurately predicted. In line with this idea and that of counterfactuals from section 1.2.2, Hitchcock (2001) equates the veracity of a SCM with the truthfulness of a family of counterfactuals that can be derived from it:

> A system of structural equations is an elegant means for representing a whole family of counterfactuals...The correctness of a set of structural equations...depends upon the truth of these counterfactuals.

1.5 Bayesian Networks (BNs)

1.5.1 Deriving BNs

In section 1.5, we will take a hiatus from our discussions on causality and simply concern ourselves with probability and probabilistic ideas such as conditional independence. In section 1.6, we shall see how the ideas here relate to causality. Let $\boldsymbol{V}$ be a set $\{V_1, \dots, V_n\}$ of n-many random variables (RVs). $\boldsymbol{V}$ admits the multivariate or joint probability distribution:

$$P(\boldsymbol{V}) = P(V_1 \cap V_2 \cap \dots \cap V_n)$$

Where "V_i" is slight abuse of notation for $V_i = v_i$ for some arbitrary value of v_i. Also, by "distribution" we mean either a mass or density function (i.e. we make no assumption on the discreteness or continuity of the RVs). The general form for a joint distribution above is agnostic towards any dependencies between the RVs. Indeed, as we will investigate in concrete terms in section 1.6, the latent causal structures of reality may induce associations between RVs or entail certain probabilistic independencies. For now, suppose that we don't particularly care about elusive ideas like causality, but rather about the fact that storing and utilising a given multivariate probability distribution can be very expensive. E.g. if our variables are discrete, the size of a joint probability table grows exponentially in the number of variables, and if we wish to perform inferences, we may have to perform many sums to marginalize out irrelevant variables to infer the probability of an event of interest. Note, by *inference* or *abduction*, we refer to the task of finding the probabilities of some events, given knowledge of some other events, i.e. evaluating one or many expressions of the form "$P(X|Y)$" (the probability of X conditioned on Y). E.g., we may wish to know the conditional distribution "$P(V_3|V_1 \cap V_8)$", and so we would first have to condition the distribution on the terms $V_1 = v_1$ and $V_8 = v_8$, and then marginalise out all other variables in $\boldsymbol{V}/\{V_1, V_3, V_8\}$ by summing. This can all be costly, and so we may look for a more compact representation of $P(\boldsymbol{V})$. An approach we may attempt is to apply the chain rule of probabilities:

$$P(\boldsymbol{V}) = P(V_1 \cap \dots \cap V_n) = P(V_1) \prod_{i=2}^{i=n} P\left(V_i \middle| \cap_{j=1}^{j=i} V_j\right)$$

Such a recursive factorisation of $P(\boldsymbol{V})$ in terms of conditional probabilities is called a Bayesian Network (BN), and we may derive from any BN, a corresponding DAG. However, such a representation does not automatically yield benefits. Indeed, the BN above is still just as costly to store and to use for inference (this can be easily checked). However, major savings may result if it turns out that for a given V_i, not all the predecessor variables in $\boldsymbol{V}$ (with respect to the given ordering/indexing of the variables) are relevant for determining the conditional distribution of V_i. Indeed, by the finiteness of the number of variables in $\boldsymbol{V}$, there will exist a minimum number of preceding variables that make the conditional distribution of V_i

independent of all other preceding variables. As an example, consider the conditional distribution "$\mathrm{P}(V_5|V_1 \cap V_2 \cap V_3 \cap V_4)$". It may very well be true that conditioned on $V_2 \cap V_3$, V_5 and "$V_1 \cap V_4$" are made independent, and so there is no need to condition on V_1 and V_4, as by the definition of conditional independence, we have that:

$$\mathrm{P}(V_5|V_1 \cap V_2 \cap V_3 \cap V_4) = \mathrm{P}(V_5|V_2 \cap V_3)$$

By repeating the process of finding *minimal* subsets of preceding variables that render a given V_i conditionally independent (CI) of all other preceding variables for every i, we obtain a BN that satisfies the so called *minimality condition*. What's more, for each V_i, the minimal set of variables that is found, may be denoted by the set $PA_i \subset \{V_1, \dots, V_{n-1}\}$ and be called the *Markovian parents* of V_i. Now, it is worth noting that for any given $\boldsymbol{V}$ (except in the degenerate case of a univariate $\boldsymbol{V}$ with a single variable), there will be distinct valid BNs with respect to every possible ordering of the random variables of $\boldsymbol{V}$. Therefore, one should not naïvely read "too much" into the DAG structure of a given BN (for example, by assuming that an arrow between 2 variables entails that there is a causal mechanism / pathway of influence between them), as indeed for the same $\boldsymbol{V}$ but with respect to a different ordering of the variables, the resulting BN may not even share the same *skeleton*, that is the set of all graphical adjacencies present in its DAG (Pearl, 2009). Furthermore, even granted a specific variable ordering, the existence of a *single* valid BN is not always guaranteed. Consider the expression "$\mathrm{P}(V_5|V_1 \cap V_2 \cap V_3 \cap V_4)$". Suppose we find that conditioning on $V_2 \cap V_3$ makes the probabilities of V_5 and $V_1 \cap V_4$ independent. However, suppose that conditioning on $V_1 \cap V_3$ also makes V_5 independent of $V_2 \cap V_4$. Then, there are 2 equally valid (both minimal in the number of conditioning variables) ways to express the conditional probability for V_5:

$$\mathrm{P}(V_5|V_1 \cap V_2 \cap V_3 \cap V_4) = \mathrm{P}(V_5|V_2 \cap V_3) = \mathrm{P}(V_5|V_1 \cap V_3)$$

In turn, choosing either term yields a different BN and DAG. Indeed, the uniqueness of a BN, up to a variable ordering, is only guaranteed once we assume that the so-called *intersection property* (Peters, 2014) holds. However, for the sake of brevity, we will not discuss this. Now, to see the utility of BNs, consider the extreme case where all the variables in $\boldsymbol{V}$ are jointly independent, that is $P(\boldsymbol{V}) = P(V_1)P(V_2) \dots P(V_n)$. In other words, every variable is independent of all its predecessors. In this extreme case, there exists only a single valid BN and DAG (a fully disconnected graph with no edges/arrows whatsoever). Clearly, we can make immense savings on the storage of the joint probability distribution, since we just need to record all marginal distributions, and so the storage requirements scale linearly with the number of variables in $\boldsymbol{V}$. Inference is also trivial, since every conditional probability is just the marginal probability of the event of interest. In general, the sparser the DAG that we obtain for a given $\boldsymbol{V}$, the greater the savings, and the easier it will be to exploit the corresponding BN.

Finally, one important matter that we have sidelined, is how the existence of a conditional independence can be verified. In practice, there is no such thing as an independence statement *oracle*, and statistical independence testing will have to be performed on data to either reject or accept an independence claim, and even if we are granted near infinite data, i.e. if we "know" $P(\boldsymbol{V})$, independence testing itself will still be technically complicated and computationally intensive. Indeed, in real world applications with limited data, if we condition on too many variables (e.g. test whether V_7 and V_6 are independent conditioned on $V_1 \cap V_2 \cap V_3 \cap V_4 \cap V_5$), then, via the curse of dimensionality, we may often lack enough samples to perform any tests (Dawid, 1979; Zhang et al., 2012; Herwartz and Maxand, 2020).

1.5.2 Semi-graphoids

Suppose that for a given $\boldsymbol{V}$, and with respect some given ordering of its n-many variables, we are granted a valid BN. Now, we may wish to know if there are any "extra" independence facts that are entailed by the set of independence facts that constitute the DAG of the BN without resorting to any statistical independence testing if we are working with data, or, for the matter, without resorting to any calculations over probabilities, in case we know the joint distribution P($\boldsymbol{V}$). To clarify what we mean, note that the DAG of a given BN may be represented as a variable ordering and a set of precisely n-many *independence statements* or *facts* of the form:

$$V_i \perp\!\!\!\perp \{V_1, V_2, \dots, V_{i-2}, V_{i-1}\}/PA_i \mid PA_i, \forall i \in \{1,2, \dots n\}$$

Where the statement "$A \perp\!\!\!\perp \mathrm{B} \mid \mathrm{C}$" means that the variables A and B are independent conditioned on C, so in the case above, we have that V_i is independent of the variables preceding it that are not its Markovian parents (i.e. "$\{V_1, V_2, \dots, V_{i-2}, V_{i-1}\}/PA_i$"), conditioned on its Markovian parents. Such a set of n-many independence statements may be referred to as a *recursive basis* (Geiger, Verma and Pearl, 1990). Note, every DAG (with a given variable ordering) describes a unique recursive basis, and every recursive basis describes a unique DAG. So, DAGs and recursive bases are equivalent in that they may be put into a 1-to-1 correspondence. What we wish to know then, is given a recursive basis / a DAG, are there any additional independence statements that are entailed (i.e. that must hold), and that we perhaps can also deduce?

As it stands, we have no tools to answer such a question, as we do not know of any operations that can be performed on the elements of a recursive basis that would allow us to derive new independence statements. We may gain such a toolkit by returning to the setting of probability theory, and seeing if we can prove any useful general rules about what additional independence statements are entailed (i.e. must be true) given some independence statement/s. For example, there is the so-called *decomposition axiom* stating that $A \perp\!\!\!\perp \mathrm{B} \cap \mathrm{C} \mid \mathrm{D} \Longrightarrow A \perp\!\!\!\perp \mathrm{B} \mid \mathrm{D}$. Note, in the context of probability theory, this "axiom" is in fact easily provable:

$$A \perp\!\!\!\perp \mathrm{B} \cap \mathrm{C} \mid \mathrm{D} \Leftrightarrow \mathrm{P}(\mathrm{A} \cap B \cap C|D) = \mathrm{P}(\mathrm{A}|D)\mathrm{P}(B \cap C|D) \;\; \forall D \text{ s.t. } \mathrm{P}(\mathrm{D}) > 0$$

$$\sum_C P(\mathrm{A}|D)\mathrm{P}(B \cap C|D) = \mathrm{P}(\mathrm{A}|D)\mathrm{P}(B|D) = \sum_C \mathrm{P}(\mathrm{A} \cap B \cap C|D) = \mathrm{P}(\mathrm{A} \cap B|D)$$

$$\therefore \mathrm{P}(\mathrm{A}|D)\mathrm{P}(B|D) = \mathrm{P}(\mathrm{A} \cap B|D), \forall D \text{ s.t. } \mathrm{P}(\mathrm{D}) > 0 \Leftrightarrow A \perp\!\!\!\perp \mathrm{B} \mid \mathrm{D}$$

We may interpret the *decomposition axiom* as a rule for generating new independence facts. For example, if our recursive basis contains the statement "$V_5 \perp\!\!\!\perp V_1 \cap V_4 | V_2 \cap V_3$", the decomposition axiom allows us to derive "$V_5 \perp\!\!\!\perp V_1 | V_2 \cap V_3$" and "$V_5 \perp\!\!\!\perp V_4 | V_2 \cap V_3$".

Importantly, the proof of the decomposition axiom did not rely on any special assumptions about the multivariate probability distribution over the variables A, B, C and D (other than assuming that "$\mathrm{P}(\mathrm{D}) > 0$", yet this is not an issue, because the initial independence statement "$A \perp\!\!\!\perp \mathrm{B} \cap \mathrm{C} \mid \mathrm{D}$" itself only applies in a context where D has been conditioned on, and we can only condition on a value of D that has a positive probability/non-zero density). So, due to the generality of its proof, the independence statements derived by applying the axiom will hold for *any* distribution $\boldsymbol{V}$. Now, without proof, we present the following 4 rules called the *semi-graphoid axioms*, all of which be proven without any special assumptions on $\boldsymbol{V}$:

$$\text{Symmetry: } A \perp\!\!\!\perp \mathrm{B} \mid \mathrm{D} \Longrightarrow B \perp\!\!\!\perp \mathrm{A} \mid \mathrm{D}$$

$$\text{Decomposition: } A \perp\!\!\!\perp \mathrm{B} \cap \mathrm{C} \mid \mathrm{D} \Longrightarrow A \perp\!\!\!\perp \mathrm{B} \mid \mathrm{D}$$

$$\text{Weak union: } A \perp\!\!\!\perp \mathrm{B} \cap \mathrm{C} \mid \mathrm{D} \Longrightarrow A \perp\!\!\!\perp \mathrm{B} \mid \mathrm{C} \cap D$$

$$\text{Contraction: } (A \perp\!\!\!\perp \mathrm{B} \mid \mathrm{D}) \wedge (A \perp\!\!\!\perp \mathrm{C} \mid \mathrm{B} \cap \mathrm{D}) \Longrightarrow A \perp\!\!\!\perp \mathrm{B} \cap \mathrm{C} \mid \mathrm{D}$$

That is, all of these "axioms" can be proven by elementary means from the standard rules of probability (indeed, we gave the proof for the "decomposition axiom") (Spohn, 1980). Now granted a recursive basis K, we may consider its *closure* under these rules, i.e. the set of all independence statements that may be derived from the recursive basis via a finite number of applications of these rules/axioms. We call the resulting set of obtainable statements the *semi-graphoid induced* by K. Again, it follows that any DAG (with a variable ordering), and by extension any BN that agrees with the recursive basis K must obey every independence statement in the semi-graphoid induced by K (because the rules/axioms themselves are provably valid, and so they can only entail true consequents when their antecedents are true). Now, setting aside for a moment any issues related to the practicality of actually finding the closure via the *semi-graphoid* axioms, we may ask ourselves whether there are any more independence statements that *must* hold true (granted some recursive basis)? That is, given just the structure/DAG of a BN, which can be described by some recursive basis, are there any more independence statements that are entailed, but which are not elements of the semi-graphoid induced by the recursive basis? Perhaps if we proved some additional "axioms/rules", beyond the 4 semi-graphoid axioms (using the rules of probability), we could gain more tools with which we could prove even more statements? As it turns out, *all* of the independence statements that are entailed by a recursive basis K are found in the semi-graphoid induced by K. In this sense, the semi-graphoid axioms are *complete*, granted they are applied to a recursive basis (Geiger and Pearl, 1990). That means there are no independence statements or facts that must hold true granted some K, which are not already elements of a semi-graphoid induced by that K. The elegant proof of this result is essentially a proof by contradiction, or rather a proof by counterexample, where it is shown that for any independence statement that is not in a given semi-graphoid induced by some K, there will always exist some joint probability distribution $P(\boldsymbol{V})$ that yields the recursive basis K, but which violates that given independence statement. A presentation of the proof itself lies beyond the scope of this thesis but can be read in Geiger and Pearl (1990). Still, let us clarify why finding the counterexample described matters at all. Again, the semi-graphoid induced by K is simply a (finite) set of independence statements that must evaluate to be true for *any* $P(\boldsymbol{V})$ that yields a BN that has a DAG that is equivalent to K. Therefore, if the semi-graphoid induced by K is missing some independence statements that hold for all possible $P(\boldsymbol{V})$ that factorise according to K, then those independence statements, which are not elements of the semi-graphoid, must never be violable. However, the proof establishes that any independence statement outside of those in the semi-graphoid (induced by K) will be violated by *some* BN that factorises according to K, as a method for constructing such a violation/BN is shown. In summary, just by looking at the structure of a BN, or equivalently, at its recursive basis, there are a set of independence statements we may derive by applying certain rules called the semi-graphoid axioms. The obtained set of independences will, logically, also apply to any other BN that share the same DAG/structure, since the independences are derived using only structural information. However, any given BN may satisfy more independencies than those

that are implied by its structure; but these will be specific to it and cannot hold for *all* BNs with that structure. So, to identify these additional independences, one must necessarily go beyond looking at the DAG structure of a given BN and must assess the actual probabilities.

Moreover, there is a condition called *faithfulness* that a BN may or may not satisfy, which states that the independence statements that hold for the BN are only those that are entailed by its structure/DAG. That is, granted a faithful BN, every true independence statement must be an element of the semi-graphoid induced by the faithful BN's recursive basis. In other words, when a semi-graphoid gives a complete specification of all the independence facts that hold for some BN, we can say that the BN is faithful, or that the BN is faithful to its DAG.

1.5.3 The d-separation criterion

Now, as powerful as the semi-graphoid axioms may seem, in practice they may not be very useful. For example, given a recursive basis K, how do we decide if we have found all the independence facts that it entails, and that there are no additional independences that can be derived if e.g. we just were to apply the right combination of the axioms to some known independence statements? Is it in theory even possible to identify if all statements have been found? Furthermore, suppose we have found the semi-graphoid induced by some K; will it be practical to store it? In the extreme case where $\boldsymbol{V}$ consists of jointly independent variables, the semi-graphoid grows *at least* exponentially in the number of variables in $\boldsymbol{V}$.

It turns out there exists an intuitive graphical criterion called *d-separation*, which runs in linear-time on the number of variables in $\boldsymbol{V}$ (Geiger, Verma and Pearl, 1990) for determining whether for a given BN, a given independence statement is entailed or not by virtue of the structure of the BN. This criterion spares us from having to deal with semi-graphoids directly. Indeed, d-separation can be proven to be *sound* via an inductive proof on the ordered elements of any given DAG/BN (Verma and Pearl, 1990). Soundness here means that every independence indicated by d-separation is in fact true, granted that the $P(\boldsymbol{V})$ we are dealing with actually factorises according to the BN that is found for it. What's more, d-separation can identify the same set of independence statements as can be obtained by the semi-graphoid axioms, and so it is also complete in the same sense. Therefore, if an independence statement is not confirmed by d-separation, there will exist some BN where precisely that independence statement fails to obtain. In particular, the d-separation criterion says that 2 disjoint sets of variables $\boldsymbol{A}$ and $\boldsymbol{B}$ are *d-separated* conditioned on some disjoint set $\boldsymbol{Z}$ (where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{Z} \subset \boldsymbol{V}$), if for every $A \in \boldsymbol{A}$ and $B \in \boldsymbol{B}$, all *paths* connecting A and B are *blocked* or *d-separated* by $\boldsymbol{Z}$. In turn, a path in a DAG, or in any graph for the matter, is an ordered sequence of edges, such that each edge leads to a node that is connected with the subsequent edge. A path in a DAG is blocked if it contains a fork "$n_1 \leftarrow n_2 \rightarrow n_3$" or a chain "$n_1 \rightarrow n_2 \rightarrow n_3$" such that the middle node denoted by "n_2", is in the conditioning set $\boldsymbol{Z}$, or if the path contains a collider "$n_1 \rightarrow n_2 \leftarrow n_3$" where neither the middle node, nor any of its graphical descendants are in $\boldsymbol{Z}$. Note, if a path is not blocked, it is *active* or *open*. Now, if every path connecting A and B is blocked by $\boldsymbol{Z}$, then A and B are d-separated and the independence statement $A \perp\!\!\!\perp B \mid \boldsymbol{Z}$ must be true. Indeed, the independence must hold, not only for the given BN, but for any BN which has the same given DAG or even a *different* DAG, as long as the same set $\boldsymbol{Z}$ d-separates A and B in the alternative graphical structure (Pearl, 2009). For a concrete example, consider:

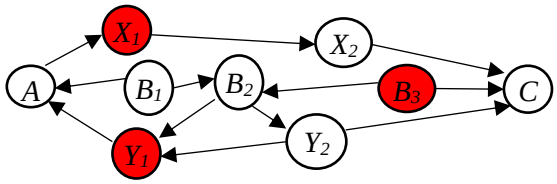


Figure 3: A DAG with a d-separating set for the variables A and C

In this DAG, $\boldsymbol{Z} = \langle X_1, Y_1, B_3 \rangle$ is a d-separating set for A and C. Note, $\langle A, B_1, B_2, B_3, C \rangle$ is a collider path, $\langle A, X_1, X_2, C \rangle$ is a directed path that is blocked by conditioning on X_1, and $\langle A, Y_1, Y_2, C \rangle$ is a confounding path blocked at Y_1. Yet, as Y_1 is a child of B_2, conditioning on Y_1 opens the collider path, hence why we also need to condition on B_3 to obtain d-separation.

Now, for any 2 distinct nodes/variables V_i and V_j in an arbitrary DAG, d-separation is always possible, except for the obvious case of when the nodes are adjacent. Indeed, assume that the nodes are not adjacent, then by extension, neither can be a parent of the other; next, let us condition on the union of the set of parents for both variables, i.e. let $\boldsymbol{Z} = PA_i \cup PA_j$. Clearly, every path that has an arrow leading into either V_i or into V_j is blocked by conditioning on $\boldsymbol{Z}$. Therefore, an open path, if it exists, must lead out from both V_i and V_j. Now, note that one of V_i and V_j will appear later in the given ordering of the variables. Also, any open path must traverse some "last" node that comes later than both V_i and V_j in the order, after which the path must move backwards in the ordering, to reach an "earlier" node. At this node, the path must be at a collider, but the collider cannot possibly be active/open because our $\boldsymbol{Z}$ only consist of nodes that are prior to the "later" node from the pair V_i and V_j. Therefore, no open path between V_i and V_j exists if we condition on $\boldsymbol{Z}$, and so the variables are d-separable (i.e. can be d-separated, as we have just shown). The result that non-adjacent nodes can always be d-separated in a DAG is impressive because it means that if 2 variables V_i and V_j of a BN are not adjacent, there exists a conditioning set such that $V_i \perp\!\!\!\perp V_j \mid \boldsymbol{Z}$ obtains.

We reemphasise that d-separation is sound: it always yields valid independence statements, but it does not necessarily uncover *all* independence statements that hold for a particular BN. Indeed, if it does, then what we have is called a faithful BN, i.e. a BN that is faithful to its DAG / i.e. a BN that satisfies the faithfulness condition. Yet, since faithfulness is not always assured, d-separation and conditional independence have their own distinct notation: "$A \perp\!\!\!\perp_G B \mid \boldsymbol{Z}$" denotes that the conditioning set $\boldsymbol{Z}$ d-separates A and B, and "$A \perp\!\!\!\perp_P B \mid \boldsymbol{Z}$" denotes that A and B are probabilistically independent conditioned on $\boldsymbol{Z}$ (Zanga and Stella, 2022). So, the soundness of d-separation can be restated as: $A \perp\!\!\!\perp_G B \mid \boldsymbol{Z} \Longrightarrow A \perp\!\!\!\perp_P B \mid \boldsymbol{Z}$. This statement is sometimes also called the *global Markov property* (Sadeghi, 2017) for BNs. The converse is the faithfulness condition, i.e. that a given BN is faithful to its DAG: $A \perp\!\!\!\perp_p B \mid \boldsymbol{Z} \Longrightarrow A \perp\!\!\!\perp_G B \mid \boldsymbol{Z}$**.** So, we have reviewed d-separation and noted it as a powerful tool for studying the independencies of any given $P(\boldsymbol{V})$ that is factorised as a BN. In turn, the independence properties of a system may be useful for performing inference or prediction: let X be a variable that is independent of another variable Y after conditioning on some $\boldsymbol{Z}$. Now, X cannot possibly be informative for a prediction of Y if $\boldsymbol{Z}$ is known. However, perhaps a greater use of BNs and d-separation lies not in any applications to prediction or inference, but goes, in a sense beyond statistics, for these formalisms can help us discover causation.

1.5.4 Faithfulness violations (FVs) in discrete BNs

Before returning to the setting of SCMs, we must briefly discuss an important theorem by Meek (1995) that our main results from this thesis critique. Consider multivariate probability distributions $P(\boldsymbol{V})$ that are discrete and finite, in the sense that for each $V_i \in \boldsymbol{V}$, the support of the variable is a finite subset of $\mathbb{R}$. Then, with respect to such distributions, we may derive so-called *discrete* BNs. Naturally, all the results for general BNs extend to this class, including the soundness and completeness of d-separation as a rule for determining independences. Now, we may ask ourselves, how *common* the situation is that an arbitrary BN features *extraneous* independences that are not entailed by d-separation? That is, how common are BNs that violate the faithfulness condition? To answer questions of this form about the frequency of some phenomena, it is necessary to define a scope, and in case we restrict ourselves to discrete BNs, the answer, as shown by Meek, it that out of all possible discrete BNs, the "proportion" (or to be precise, the *measure*, for any measure dominated by the Lebesgue measure, such as any probability density over an appropriately defined parameter space) that feature *any* faithfulness violation is 0. It follows that *almost all* discrete BNs are faithful, i.e. the only probabilistic independence statements that are true are the structurally entailed independences entailed by d-separation. The details of Meek's proof lie beyond the scope of this thesis but, at a high level, the idea is to find a way to parametrise discrete BNs, so that every possible discrete BN corresponds with a unique finite set of real number parameters. This is achieved by exploiting the fact that every variable $V_i \in \boldsymbol{V}$ can only assume finitely many possible values, and that the set of its Markovian parents PA_i, can also only range over a finite set of possible joint values. Then, for every given value of the parents $PA_i = pa_i$, the conditional distribution $P(V_i|PA_i = pa_i)$ can be specified by $r_i - 1$ many real numbers, where $r_i = |\mathrm{supp}(V_i)|$ (i.e. where the support of V_i is the set $\mathrm{supp}(V_i) = \{x \in \mathbb{R}: P(V_i := x) > 0\}$) with each parameter $\theta^x_{pa_i}$ specifying the probability that V_i will evaluate to some $x \in \mathrm{supp}(V_i)$. Furthermore, by the law of total probability, it follows that $\sum_{x=1}^{x=r_i} \theta^x_{pa_i} = 1$. Next, Meek shows that every independence statement, e.g. $A \perp\!\!\!\perp_p B \mid \boldsymbol{Z}$ for disjoint $A, B, \boldsymbol{Z} \subset \boldsymbol{V}$ corresponds to a finite number of polynomial equations over the BN's parameters, by exploiting the fact that any independence statement is essentially a collection of equality constraints, e.g:

$$A \perp\!\!\!\perp_p B \Leftrightarrow \forall a \in \mathrm{supp}(A)\ \forall b \in \mathrm{supp}(B),\ P(A = a \cap B = a) = P(A = a)\,P(B = a).$$

Then, it is shown that at least some of these polynomial equations are non-trivial (the corresponding polynomials are not identically 0) precisely when the conditional independence is not entailed by d-separation, by presenting a process for deriving a parametrisation where at least some of the equations are violated. Finally, a fairly elementary result from measure theory states that non-trivial polynomials over the real numbers have solution sets of Lebesgue measure 0 (Okamoto, 1973). It follows that the sets of parameters that yield discrete BNs with an exogenous independence statement (i.e. an independence statement not entailed by d-separation) have measure 0 over all possible parameterisations. Now since the number of possible independence statements themselves are finite (for a BN with finitely many variables) it follows, by the fact that finite unions of sets with measure 0 have measure 0, that the set of parameterisations for discrete BNs that yield *any* extraneous independence statement has measure 0. Meek's conclusion then is, if we were to consider an arbitrary discrete BN, it will almost surely (with probability 1) be faithful to its DAG. As we will further discuss in chapter 2, our main critique of this theorem boils down to how unconstrained and large the set of *all*

possible discrete BNs is, and that many of the discrete BNs that we may wish to consider will come from much more constrained subclasses that are in fact so small (with Lebesgue measure 0), that faithfulness violations can occur with non-0 measure, i.e. a positive probability.

1.6 Non-deterministic SCMs and causal BNs

1.6.1 Non-deterministic Markovian SCMs (MCMs) and causal BNs

We noted in section 1.4.2 that granted the observability of the exogenous variables $\boldsymbol{U}$ and the acyclicity of the causal mechanisms modelled by an SCM, the values of the endogenous variables $\boldsymbol{V}$ become uniquely determinable and can be computed via $\boldsymbol{f}$. Indeed, we could not hope for a more ideal scenario. In general, there will be limits to our knowledge on the pertinent events of a system and on the conditions outside of the system that drive the causal mechanisms determining the values of our measurements for $\boldsymbol{V}$, i.e. on the set $\boldsymbol{U}$. Hence, we drop assumption that $\boldsymbol{U}$ is observable, and instead interpret $\boldsymbol{U}$ as a set of random variables (RVs) obeying some joint distribution $P(\boldsymbol{U})$, the shape of which we may not know, yet which, unless indicated otherwise, we assume to be stable. We invoke RVs because the characteristics and behaviours of systems that are not mechanistically/causally well understood and their surroundings, can be modelled probabilistically, i.e. we may not be able to explain, via any causal mechanisms *why* some events obtain, but we can still describe and predict them. Note, probabilities, in theory *could* permit us to model inherently non-deterministic systems (e.g. involving quantum mechanical processes), but our assumption is that we are working in deterministic settings, and so we interpret their central utility as a means to deal with unknown boundary conditions. Indeed, assuming acyclicity and no interventions, any $P(\boldsymbol{U})$ induces, via $\boldsymbol{f}$, a unique *observational distribution* $P(\boldsymbol{V})$ on $\boldsymbol{V}$. Furthermore, instead of keeping the functional notation "$V_i := f_i(PA_i, U_i)$", we may use the conditional probability expression "$P(V_i|\ PA_i)$" for V_i conditioned on its observable parent variables PA_i, where "$P(V_i|\ PA_i)$" is slight abuse of notation for $P(V_i = v_i|\ PA_i = pa_i)$ for some arbitrary values of v_i and pa_i. Note that U_i does not feature in the conditional probability expression, as it now is unobservable. However, is this permissible? Now, that every given U_i is indeterminate, surely there is a possibility that measuring some $V_j \notin PA_i$ let's us infer something about U_j, which by virtue of some possible latent causal mechanisms, may let us infer something about U_i, and so change our estimate for "$P(V_i|\ PA_i)$". This scenario is problematic due to various reasons. Among them, is that it negates the idea of the conditional distributions "$P(V_i|\ PA_i)$" from serving as representations of stable and conserved causal mechanisms of the system, if indeed the form of each such expression can vary whenever we make some further observations.

To handle this issue, we introduce the substantial, yet common assumption that an SCM is *complete* or *sufficient* (Spirtes et al., 2001, p. 27) meaning that all the variables of $\boldsymbol{U}$ are taken to be mutually independent: $P(\boldsymbol{U}) = \prod_{i=1}^{i=n} P(U_i)$. This can be a very substantial assumption, and in a sense, amounts to the claim that the variables in $\boldsymbol{V}$ are so well chosen and so informative, that any unaccounted/unmeasured events and causal mechanisms that influence $\boldsymbol{V}$ are so specific, localised and weak in their scope of influence that they fail to confound any groups of endogenous variables. All these requirements may be violated. Therefore, additional

arguments are often made in theoretical support of causal sufficiency. According to some theorists such as Judea Pearl, we, in a sense, only begin to understand systems once sufficiency can be assumed, at least in part because its failure, i.e. the existence of significant latent causal mechanisms that generate dependencies among the variables in $\boldsymbol{U}$ amounts to the idea that we are measuring a system too superficially to be able to model its "inner" functioning, and in practice, we will have to keep on exploring and acquiring more variables and measurements, until essentially, all would-be confounding events can be measured by a sufficiently expanded set of endogenous variables, and that any remaining unaccounted sources of influence have been rendered disparate and jointly independent. Others, such as Spirtes, Scheines and Glymour have justified the condition because it appears to be empirically satisfied in many real-world examples, and that under certain circumstances, e.g. when some variable undergoes experimental randomisation, the principle can be expected to hold, or rather, *must* be expected to hold in order to theoretically explain how methods such as randomisation can work at all (Hitchcock, 2023). In any case, the assumption of sufficiency serves as a lynchpin for many methods dedicated to the study of causal systems: e.g. to the problem of *causal discovery*, which refers to the task of recovering unknown or underspecified SCMs, or at least some of their important aspects, e.g. their causal structure/DAG, given only *observational* data, i.e. data that does not include any interventions and their outcomes. Although, let it be said that causal discovery under causally *insufficient* settings also has received a great amount of research attention and effort (Glymour, Zhang and Spirtes, 2019).

In any case, for the remainder of this thesis, we will keep with the simpler option, and assume causal sufficiency/completeness. SCMs that are both sufficient and acyclic are called *Markovian* SCMs or MCMs in short, for which one may derive a fundamental condition called the causal Markov (CM) condition (Hausman and Woodward, 1999). Note, that the acyclicity (and finiteness) of an MCM allow us to define a *causal order* that respects all causal kinship relations, i.e. if V_j is a causal ancestor of V_i, then $j < i$ holds. Now, with respect to a causal order, and given some $V_i \in \boldsymbol{V}$, let $R_i = \{V_1, \dots, V_{i-1}\}$ be the set of V_i's predecessors. It follows that $PA_i \subset R_i$, and that $R_i / PA_i \perp\!\!\!\perp_P V_i \mid PA_i$. This independence is guaranteed for MCMs, as conditioned on some PA_i, the only variable that may still influence the value of V_i is U_i, and by our assumption of sufficiency, $U_i \perp\!\!\!\perp_P \boldsymbol{U} \backslash U_i$, and so in particular: $U_i \perp\!\!\!\perp_P \{U_1, \dots, U_{i-1}\}$. In turn, by the acyclicity assumption, every $V_j \in R_i$ is a function over $\{U_1, \dots, U_{i-1}\}$, and it follows by the fact that functions of independent variables are independent, that for every $V_j \in R_i$ that $V_j \perp\!\!\!\perp_P U_i$ is satisfied. In turn, applying the weak union axiom yields $R_i \backslash PA_i \perp\!\!\!\perp_P U_i \mid PA_i$, and as stated, conditioned on PA_i, V_i is a function of U_i, so we obtain the desired result that $R_i \backslash PA_i \perp\!\!\!\perp_P V_i \mid PA_i$. Amazingly, this statement means that we may interpret any MCM as a BN, or rather a so-called *causal* BN in line with its causal DAG.

Furthermore, for a given V_i, since every causal parent variable in PA_i is presumably measuring some causes or core for some atomic mechanism that can determine the value of V_i, each $V_j \in PA_i$ should be functionally relevant in f_i, and by extension, we expect to observe statistical dependencies of the form $V_j \not\perp\!\!\!\perp_P V_i \mid PA_i \backslash V_j$. The intuition is that if some V_j measures even a single event (with a positive probability) that, under some core, features as a cause in an atomic mechanism that induces a value for V_i, then granted such a core, which can presumably be done by conditioning on the right values for "$PA_i \backslash V_j$", the causal dependence will manifest as a statistical dependence. Indeed, if for some causal BN, all such expected

dependencies obtain, we say that that the causal BN satisfies the *causal* minimality (CMin) condition. As a counterexample, consider an MCM with $V_3 := f_3(V_1, V_2, U_3) = V_1 + V_2 + U_3$, where $V_2 := f_2(V_1, U_2) = V_1$ and $V_1 := f_1(U_1) = U_1$. Now, the causal BN directly derived from these structural equations would include the factor "$\mathrm{P}(V_3|V_1 \cap V_2)$", even though one of the parents appears to be superfluous, since knowing one endogenous parents renders the other parent redundant, e.g. $\mathrm{P}(V_3|V_1 \cap V_2) = \mathrm{P}(V_3|V_1)$. Now, if we intervened in the system, e.g. to override the mechanism "$f_2(V_1, U_2) = V_1$", the conditional independence "$V_3 \perp\!\!\!\perp_P V_2 \,|V_1$" could be violated, and the expression "$\mathrm{P}(V_3|V_1)$" could cease being accurate. In light of this, we prefer a robust non-minimal causal BN that reflects the underlying causal structure of an SCM to a BN that is minimal but doesn't. However, note any tolerated violations of the CMin condition preclude a causal BN from being faithful. E.g. V_3 and V_1 being graphically adjacent means that they cannot be d-separated, and so the idiosyncratic conditional independence $V_3 \perp\!\!\!\perp_P V_2 \mid V_1$ does not have a corresponding d-separation $V_3 \not\perp\!\!\!\perp_G V_2 \mid V_1$. Hence, we have an independence that d-separation cannot detect. So, causal BNs in general need not be faithful to their causal DAGs, that is, satisfy the so-called causal faithfulness condition (FC). On this note, we arrive at our final background section on the topic of faithfulness violations (FVs).

1.6.2 FVs in MCMs

We have just shown how a causal BN can violate FC by failing the CMin condition. Conversely, if we assume FC, then any 2 nodes A and B that are adjacent in a causal DAG cannot be rendered conditionally independent (CI) as adjacent nodes in general cannot be d-separated (as the path corresponding to the edge between the nodes cannot be blocked). So assuming FC, every conditional independence must correspond to a d-separation, and finding a way to make 2 adjacent nodes CI would correspond to finding a way to d-separate adjacent nodes, which is impossible. In contrast, we showed in section 1.5.3 that any non-adjacent pairs of nodes in a DAG can always be d-separated, and so may always be rendered CI. This distinction between adjacent nodes, which cannot be rendered CI (granted FC), and non-adjacent nodes, which can always be d-separated, and hence rendered CI, serves as the basis of a whole class of causal discovery algorithms known as *constraint-based* methods (Glymour, Zhang and Spirtes, 2019), including the influential PC algorithm (Spirtes and Glymour, 1991). The core idea of such algorithms is that, if after some form of an intelligently performed search, no conditioning set $\boldsymbol{Z}$ is found that renders a given pair of variables A and B CI, then the variables must be adjacent (otherwise there would *have* to exist some d-separating set, which we eventually would find). Hence, as the only alternative, A and B must in fact be adjacent. To be clear, if a faithful causal BN is derived from an MCM, then showing the adjacency of any 2 nodes in the causal DAG amounts to showing that one of the corresponding variables must be measuring events that cause events measured by the other variable, since latent common causes are precluded by our assumption of causal sufficiency. So granted an MCM with a faithful BN, we can infer and rule out the existence of some causal dependencies/mechanisms just via observational data obtained from the MCM. Now, a way to try assessing the viability of causal discovery methods that hinge upon FC is to theoretically investigate how common FVs are, as it can easily be shown that methods that assume FC may fail when there are FVs. Note, we already briefly discussed in section 1.5.4, Meek's proof that for discrete BNs, and by extension, over *discrete MCMs*, the proportion of systems that will

present FVs will presumably be measure 0, making the FC assumption seem reasonable to hold in practice if we are working with systems with discrete variables. Yet before we move on to criticising this view in chapter 2, as a final topic of review, let us briefly discuss some of the literature on FVs. Broadly speaking, there are 2 categories of FVs that are commonly considered in the literature: FVs due to *path cancellation* and FVs due to *determinism*.

Path cancellation occurs when different pathways of causal influence precisely negate each other. Consider the following MCM with 3 endogenous variables: $V_1 := U_1$; $V_2 := aV_1 + U_2$; $V_3 := U_3 + bV_2 + cV_1$. If the equation $ab + c = 0$ obtains, then V_1 and V_3 will be marginally independent as $U_3 + bV_2 + cV_1 = U_3 + b(aV_1 + U_2) + cV_1 = U_3 + bU_2 \perp\!\!\!\perp_P V_1$, that is despite the direct causal dependence of V_3 on V_1. So, conditioned on the empty set, V_1 and V_3 are d-connected, as there are 2 open directed paths from V_1 to V_3, yet the effects of these paths precisely cancel out, and the independence of V_1 and V_3 follows from sufficiency (the mutual independence of U_1, U_2 and U_3). Now, in the context of SEMs, and in general, for MCMs where an assumption on the linearity of the functions in $\boldsymbol{f}$ is made (like above), the FC condition has been justified via a line of argumentation that is similar to Meek's result for discrete BNs, namely by proving that violations may only occur over "fine-tuned" parameters that solve certain algebraic equations (e.g. $ab + c = 0$) and which occupy, over appropriately defined parameter spaces, sets of Lebesgue measure zero (Spirtes et al., 2001, p. 41). In turn, there is a counterargument that faced with limited sample sizes in the real world, systems with "near-FVs" due to "near" path cancellations, which encompass parameter sets with non-zero measure, may also induce false negatives in causal discovery algorithms, i.e. incorrect judgements about non-causation (Uhler et al., 2013). Indeed, as we have already mentioned, limitations in sample size can present a major issue for causal inference in general (Spirtes et al., 2001, p. 77). Furthermore, some have argued that the commonness of systems (of interest) that violate faithfulness may be greater than what measure-theoretic arguments suggest (Andersen, 2013); e.g. systems that assume steady states via regulatory mechanisms that negate disturbances, such as living organisms maintaining various forms of homeostasis, can violate faithfulness (Weinberger, 2023). As an example, consider the core body temperature of a person and the ambient temperature. If the latter varies within some well tolerable range (e.g. 15-30 C), we may find that core body temperature independently keeps near 36-37 C, even though the ambient temperature causally influences the body by drawing out heat. However, such heat loss is compensated for by homeostatic processes, possibly yielding a FV.

Moving on, the example of an FV that we gave at the end of 1.6.1 hinged upon V_2 being a deterministic function of V_1. The presence of determinism, or in general of *information equivalence* (Lemeire, Steenhaut and Maes, 2012) between 2 disjoint sets of variables $\boldsymbol{X}, \boldsymbol{Y} \subset \boldsymbol{V}$ means that conditioning on one set, e.g. $\boldsymbol{X}$, allows us to determine the value for the other set $\boldsymbol{Y}$, and since determined/constant variables are CI w.r.t any other variables including any adjacent ones, conditioning on e.g. $\boldsymbol{X}$ can yield many FVs. Accordingly, alternative criteria to d-separation have been suggested for deterministic nodes (Spirtes et al., 2001, pp. 53-58). However, there appear to be relatively few causal discovery methods that handle determinism-induced FVs (Mabrouk et al., 2014; Shajarisales et al., 2015; Yang et al., 2022; Zeng et al., 2023; Li et al., 2023). This is perhaps because determinism has been considered trivial to detect in practice (Steel, 2006), or perhaps due to it being a direct conduit for FVs, there being an assumption that its prevalence must be negligible by arguments such as Meek's, which in a blanketed manner, suggest that MCMs with *any* form of FVs should be very rare.

Chapter 2

Results and Discussion

2.1 Discrete MCMs with discrete noise

Even if a system is fundamentally not discrete, i.e. its number of conceivable states are infinite, it may be inevitable to take a discrete *view* of the system, as in practice the resolution of any measurement instrument will be limited and thereby discrete. So, if we wish to theoretically assess how common FVs are across both continuous and discrete systems, it may suffice, at least for practical considerations, to settle the question for discrete systems. In turn, recall that the prevalence of FVs matters because they present issues for causal discovery (e.g. induce spurious causal independencies). One of the simplest scenarios is when we assume both causal sufficiency and acyclicity, i.e. when our set of measurement/endogenous variables and the causal mechanisms for a given system correspond with a discrete MCM, i.e. a MCM where $\forall V \in \boldsymbol{V}$, $|\mathrm{supp}(\mathrm{V})| < \infty$. Now, recall Meek's result (1995) that over discrete BNs, which may be interpreted as causal BNs for discrete MCMs, FVs occur with a Lebesgue measure, and by extension, with a probability measure of 0 (granted a probability density function over the space of possible parameters). It may appear that the matter is settled. However, we believe that the general class of discrete BNs, which Meek parameterises, is *too general*. Recall that the probability for a variable $V_i \in \boldsymbol{V}$, conditioned on its parents has the form $\theta^{x}_{pa_i} = \mathrm{P}(V_i = x|PA_i = pa_i)$, where $\sum_{x=1}^{x=r_i} \theta^{x}_{pa_i} = 1$, $\mathrm{supp}(V_i) = r_i < \infty$, and $\mathrm{supp}(PA_i) < \infty$. By the non-negativity of probabilities, it also follows that $\forall x, 0 \leq \theta^{x}_{pa_i} \leq 1$. Now, note an arguably *loaded* assumption that Meek makes: namely, that the parameters $\{\theta^{1}_{pa_i}, \dots, \theta^{r_i-1}_{pa_i}, 1 - \theta^{1}_{pa_i} - \cdots - \theta^{r_i-1}_{pa_i}\}$ are sampled from a joint probability *density* (in $r_i - 1$ dimensions), since Meek assumes that the measure according to which the parameters are sampled is dominated by the Lebesgue measure. That is, if we let λ be the Lebesgue measure over $r_i - 1$ dimensional Euclidean space; then another measure ν is said to be *dominated by* λ, if and only if $\forall A \subset \mathbb{R}^{r_i-1}, \lambda(A) = 0 \Longrightarrow \nu(A) = 0$; in turn, if μ is a probability measure dominated by the Lebesgue measure, it must assign measure, and hence probability 0 to any set of Lebesgue measure 0. However, our belief is that under certain assumptions, which we will try to justify on intuitive grounds, we can expect to sample, with positive probability, model parameters from certain sets of parameterisations with Lebesgue measure 0. In turn, this violation of Meek's assumption annuls the applicability of his argument, and therefore it comes as no surprise and contradiction, that under our assumptions, FVs appear with positive probability.

As the first and simplest of our scenarios, where we may expect to see FVs with positive probability, consider the class of all discrete MCMs with *discrete exogenous variables* or *noise*. That is, assume that for $\forall U_i \in \boldsymbol{U}$, $|\mathrm{supp}(U_i)| = q_i < \infty$ obtains. We will, in short,

discuss what such an assumption amounts to. Now, by the discreteness of our MCM, we know that $\forall V_i \in \boldsymbol{V}$, $|\text{supp}(V_i)| = p_i < \infty$. In turn, it follows that every $f_i \in \boldsymbol{f}$ *must* be a mapping between finite sets, and so for any given i, the set F_i of possible specifications for "f_i" is itself finite. Now, if we randomly sample a specification for f_i from F_i according to a probability mass function over F_i that is positive (i.e. $\forall f \in F_i, p(f) > 0$ where "$p(f)$" is the probability of sampling the specification "f"), then there is a positive probability that we will sample a specification for f_i that is functionally independent of U_i, i.e. where $\exists g \text{ s.t. } f_i(PA_i, U_i) = g(PA_i)$. In turn, for any child V_j of the variable V_i (we can always find some V_i with children, if our MCM has more than 1 endogenous variable and the set $\boldsymbol{f}$ of modelled causal mechanisms is not empty), we will have the FV: $V_j \perp\!\!\!\perp_P V_i | PA_i$ that holds due to *effect determinability*, i.e. due to V_i being a deterministic function of PA_i, while there fails to be a corresponding d-separation, i.e. $V_j \not\perp\!\!\!\perp_G V_i | PA_i$. Indeed, a method such as PC, that would extrapolate from the CI "$V_j \perp\!\!\!\perp_P V_i | PA_i$" the non-adjacency of V_i and V_j would commit an error. Therefore, we have demonstrated that over the space of discrete MCMs with discrete noise and causal DAGs with at least 2 nodes and 1 arrow, the "proportion" of possible MCMs that will feature FVs is greater than 0. As we have discussed, the key to this result is that the finite set of possible mapping rules F_i from $PA_i \cap U_i$ to V_i will have a positive proportion of functions that are constant in U_i, i.e. $\exists f \in F_i \text{ s.t. } \forall pa_i \in \text{supp}(PA_i), \forall u_i^1, u_i^2 \in \text{supp}(U_i),\ f(pa_i, u_i^1) = f(pa_i, u_i^2)$.

Now, an objection to our argument is that it is contingent on sampling the $f_i \in \boldsymbol{f}$ via a distribution / probability mass function that is positive over all possible specifications (or at least, positive over *some* specification where U_i is redundant), yet with $\boldsymbol{U}$ serving, by definition, to account for *relevant* unmeasured events, considering any functions that don't depend on $\boldsymbol{U}$ seems contradictory and confused. However, for the sake of our argument, we have not taken such an empirical and post hoc view of $\boldsymbol{U}$ as *the* supplementary information about our system that lets determine what atomic causal mechanisms and measurements will be elicited. Rather, we generously and implicitly have allowed ourselves to regard $\boldsymbol{U} \cup \boldsymbol{V}$ as a generic and full description of our system, and have then asked ourselves, if we then consider an MCM over such a set $\boldsymbol{U} \cup \boldsymbol{V}$ with arbitrarily chosen $\boldsymbol{f}$, how common should FVs be? Also, our result can even be strengthened if we assume some form of ordinality in $\boldsymbol{U} \cup \boldsymbol{V}$ and discrete continuity in $\boldsymbol{f}$; i.e. if we assume some condition roughly stating that for "close" values of pa_i^1, u_i^1 and pa_i^2, u_i^2, $f_i(pa_i^1, u_i^1)$ and $f_i(pa_i^2, u_i^2)$ should be "close", though a further discussion of this lies beyond the scope of our thesis. In our view, a more substantial objection to our argument lies with our assumption that $\boldsymbol{U}$ can be regarded as discrete. Via this assumption, we are baking into our MCM the condition that, for the purposes of the mechanisms determining $\boldsymbol{V}$, the possibly enormous span of unmeasured events both from within the system and from beyond it, can be reduced to a finite collection of distinct cases. However, unless our system, and the external structures that may realistically influence it are inherently discrete (e.g. if our system is a subcomponent of a digital circuit), or unless we have additional reasons to justify why the causal mechanisms for $\boldsymbol{V}$ permit us to take a discrete and finite view over all unmeasured events (i.e. why the causal mechanics of our system are driven in accordance with discrete measurements of the world), our assumption seems questionable.

2.2 Discrete MCMs with weak noise

Instead of trying to defend the discreteness of $\boldsymbol{U}$, we may argue for its *weakness*. Recall from section 1.6.1 our discussion on sufficiency for MCMs, i.e. the assumption that there are no unmeasured confounding events/variables, so that all the variables of $\boldsymbol{U}$ may be assumed to be probabilistically mutually independent. A major justification for this assumption rests on the idea that we by sufficiently expanding the scope of our system, that is, the set of endogenous variables $\boldsymbol{V}$ that we measure (and/or by improving the resolution of measurements) can arrive at a depth and breadth of knowledge, where any unexplainable variability in the value of any $V_i \in \boldsymbol{V}$ is purely attributable to events that will not have had the capacity to have influenced any other $V_j \in \boldsymbol{V} \setminus V_i$. In other words, if there is confounding among some of the variables in $\boldsymbol{V}$, we may iteratively add any would-be confounders to $\boldsymbol{V}$ until no confounders are left.

Now, we say that a discrete MCM has a *weak* noise term U_i if $\forall V_i \in \boldsymbol{V}$ s.t. $PA_i \neq \emptyset$, $\forall pa_i \in \mathrm{Supp}(PA_i)$, $|\mathrm{Supp}(V_i|PA_i = pa_i)| \ll |\mathrm{Supp}(V_i)|$. That is, U_i is weak, if most of the variation for the corresponding endogenous variable $V_i \in \boldsymbol{V}$ is determined by a non-empty set of endogenous parents PA_i, in the sense that the conditional probability $P(V_i|PA_i)$ is only ever positive over a small subset of $\mathrm{supp}(V_i)$, that is the support of the unconditional / marginal distribution of V_i. Now, we will not attempt to prove this condition, because it is in fact an assumption that can be violated, and it is easy to find counterexamples: consider a discrete MCM with $V_1 := U_1$ and a child $V_2 := V_1 + U_2$, where $\mathrm{supp}(U_2) \gg \mathrm{supp}(U_1)$. However, we believe that weak noise can reasonably we expected in many MCMs. Our intuitive argument is: if we have measured enough variables to eliminate all sources of confounding (so as to be able to assume sufficiency), it is not far-fetched to imagine that we also have a rich enough description of our system via $\boldsymbol{V}$ that we can model most of the variation in $V_i \in \boldsymbol{V}$ via $\boldsymbol{V} \setminus V_i$; in particular, for every additional source of confounding that we account for by including an additional variable in $\boldsymbol{V}$, we also increase the proportion of the variation for each $V_i \in \boldsymbol{V}$ that we can explain via the expanded set of variables in $\boldsymbol{V} \setminus V_i$. What's more, for any remaining unmeasured sources of influence on $V_i \in \boldsymbol{V}$ to not confound, i.e. to not influence any other variables $V_j \in \boldsymbol{V} \setminus V_i$, it presumably helps for the impact of all such sources to be weak (and to be localised by virtue of their weakness), as any unmeasured event with a substantial impact on some aspect of a system measured by $V_i \in \boldsymbol{V}$, would conceivably also have a greater tendency to influence some other aspects of the system, as measured by $\boldsymbol{V} \setminus V_i$ and thereby confound. We reiterate that these have been intuitive arguments, which may be easily rebutted by trivially simple counterexamples, but the same can be said about the assumptions of acyclicity and sufficiency. Therefore, our main motivation here has not been to, in any sense, *prove* the assumption of weakness, but rather to argue that the weakness of $\boldsymbol{U}$ is not unreasonable to assume, especially if we have assumed sufficiency, as is the case in any MCM.

In addition to weakness, let us now introduce another condition called *simplicity*. A function $f_i \in \boldsymbol{f}$ for some $V_i \in \boldsymbol{V}$ is *simple with respect to an endogenous parent* $PA_i^j \in PA_i$, if $\exists g_i^j : \mathbb{R} \to \mathbb{R}$ s.t. $|\mathrm{supp}(g_i^j(PA_i^j))| \ll |\mathrm{supp}(PA_i^j)|$, and $\exists q_i^j : \mathbb{R}^{m+1} \to \mathbb{R}$ s.t.:

$$V_i := f_i(PA_i, U_i) = q_i^j(PA_i^1, \dots, PA_i^{j-1}, g_i^j(PA_i^j), PA_i^{j+1}, \dots, PA_i^{m_i}, U_i)$$

Where V_i has m_i endogenous parents in $\boldsymbol{V}$. Essentially, the simplicity of f_i with respect to PA_i^j

amounts to the assumption that the amalgam causal mechanism for determining V_i is sensitive to the values of PA_i^j at a much lower resolution than at which the variable is being measured. Differently put, if our measurements for some V_j are fine-grained enough, then many of the causal mechanisms in which V_j features as a cause may simply be insensitive to much of the fine-grained variation in V_j. Simplicity should seem like a reasonable condition, though we admit that we do not have as concise arguments to justify it, as we have had for the case of weakness. Indeed, a thorough assessment of conditions under which simplicity becomes expectable in $\boldsymbol{f}$, or of how much the assumption may be weakened for the purpose of generating FVs, lies beyond the scope of our thesis, though could be pursued in future work.

Now, for some discrete MCM, let $f_i \in \boldsymbol{f}$ be *simple* w.r.t. a parent V_j of V_i. That is: $V_i := q_i^j(PA_i^1, \dots, PA_i^{j-1}, g_i^j(V_j), PA_i^{j+1}, \dots, PA_i^{m_i}, U_i)$, where $|\text{supp}\left(g_i^j(V_j)\right)| \ll |\text{supp}(V_j)| = r_j$. Indeed, let us assume that $|\text{supp}\left(g_i^j(V_j)\right)| = 2$. Such a scenario could for example arise when the influence of V_j on V_i is mediated via some fixed threshold mechanism (i.e. a non-linear step function), e.g. such that under the threshold value, V_j fails to have any influence on V_i, and above the threshold, the "full" influence of V_j on V_i is transmitted (or rather the influence of the events that V_j measure on the events measured by V_i). In any case, thresholded or not, and regardless of whether V_j is ordinal, we assume that $\left|g_i^{j^{-1}}(l_t)\right| = |d_t^j| = |\{v_j \,|(v_j \in \text{supp}(V_j)) \wedge (g_i^j(v_j) = l_t)\}| \approx \frac{r_j}{2}$, for $l_t \in \text{supp}\left(g_i^j(V_j)\right) = \{l_1, l_2\}$ (as we have assumed that $|\text{supp}\left(g_i^j(V_j)\right)| = 2$). Now, assuming the weakness of U_j, we find that there will be a non-0 probability of choosing a function f_j s.t.:

$$\forall pa_j \in \text{supp}(PA_j)\ \exists t \in \{1,2\}\ \text{s.t.}\ \forall u_j \in U_j,\ f_j(pa_j, u_j) \in d_t^j$$

This follows, as by virtue of the weakness of U_j, the values of PA_j are assumed to determine "most" of the variation in V_j, meaning that the value of U_j only specifies V_j over a small subset of $\text{supp}(V_j)$. Explicitly, let $\sup(\{|\text{supp}(V_j|pa_j)| \mid \forall pa_j \in \text{supp}(PA_j)\}) = s_j \ll \frac{r_j}{2}$. We find, for a given $PA_j = pa_j$, that the probability that the statement $\text{supp}(V_j|pa_j) \subset d_1^j \vee \text{supp}(V_j|pa_j) \subset d_2^j$ is true for a *permissibly* and *randomly* specified mechanism f_j is bounded from below by the value $\left(|d_1^j|\right)^{s_j} + \left(|d_2^j|\right)^{s_j}$. In turn, the probability that the statement obtains for *every* $pa_i \in \text{supp}(PA_j)$ is bounded from below by $\left(\left(|d_1^j|\right)^{s_j} + \left(|d_2^j|\right)^{s_j}\right)^{|\text{supp}(PA_j)|}$, which is greater than 0. To clarify, having assumed the weakness of U_j, a *permissible* specification for f_j corresponds to any mapping rule where $\forall pa_j \in \text{supp}(PA_j),\ |\text{supp}(V_j|pa_j)| \ll |\text{supp}(V_j)|$. Explicitly, conditioned on a $pa_j \in \text{supp}(PA_j)$, the set $\{f_j(pa_j, u_j) \mid \forall u_j \in U_j\}$ must be a small subset of the unconditioned range of f_j. In turn, a *randomly* chosen permissible specification of f_j obtains, if for every $pa_j \in \text{supp}(PA_j)$, the conditional range of $f_j(pa_j, u_j)$ (where u_j may vary over the entire support of U_j) is a small *and* randomly chosen subset of the unconditional range of f_j, i.e. of $\text{supp}(V_j)$. So, with respect to each $pa_j \in \text{supp}(PA_j)$, we may define the conditional range of f_j, by sampling without replacement values from

$\text{supp}(V_j)$, until a small number of values, e.g. at most s_j many values, have been sampled; then f_j may arbitrarily map every of value of U_j to a value in this conditional range. So, granted this general procedure that respects the weakness of U_j, there is a positive probability that an f_j will be chosen that in conjunction with the simplicity of f_i **(**with respect to V_j), will yield an FV. Specifically, there is a positive probability of having a g_i^j s.t. conditioned on PA_j (recall, there are only finitely many values), $g_i^j(V_j)$ will be constant $\forall V_j \in \text{supp}(V_j|PA_j)$ due to effect determinability, making V_j and V_i CI, which, due to their adjacency, constitutes an FV.

Upon reflection, the plausibility of the simplicity can come into question, as many seeming examples of simplicity can be deceiving. We consider again the idea of threshold mechanisms via a system consisting of a fixed string with a hook and some weights. Presumably, there is some critical tension beyond which the string will snap (within a given time frame, e.g. 10 seconds), and suppose that the threshold lies at 14.5 kg. If we weigh objects at the resolution of grams, we expect that the snapping of the string is independent of whether an object weighs 13,453g, 13,479g or 14,221g as all these values are under the threshold. So, letting V_1 be the weight of an object in grams, and V_2 the "measurement" of whether our string snaps (and leaving all other events/variables unmeasured), we may expect to find that $V_2 := f_2(V_1, U_2) = q_2^1(g_2^1(V_1), U_2)$, where $g_2^1(x) = 1$ if $x \geq 14.5$; else 0. However, in practice, it will likely be the case that the exact threshold value of 14.5 kg varies with respect to unmeasured events/variables e.g. the temperature, humidity, what object string is tied to, etc. That is, g_2 itself is a function of U_2, but now, the expression "$q_2^1(g_2^1(PA_2, U_2), U_2)$" seems no better than $f_2(PA_2, U_2)$. It is beyond the scope of our thesis, but in the face of such complications, we may still salvage our argument: roughly, we must determine the range of variability for a given threshold mechanism g_i^j with respect to all pertinent variables (both endogenous and exogenous). Then, we must identify all conceivable values of the variable V_j that can fall on varying sides of the threshold for different threshold values. The set of all such values may be a *small* subset of $\text{supp}(V_j)$, and we may consider the probability of sampling a mechanism f_j that maps V_j into the complement of this set. The rest of our argument can then be applied to demonstrate that the probability of obtaining an extraneous CI is still positive.

2.3 Discretisable MCMs with weak additive noise

We will now consider *discretisable* MCMs with continuous endogenous variables $\boldsymbol{V}$ with bounded support over $\mathbb{R}$, where the exogenous variables $\boldsymbol{U}$ are assumed to be *additive* and *weak*, and where all the functions represented by $\boldsymbol{f}$, can be *discretised*:

$$V_i := f_i(PA_i, U_i) = q_i(g_i^1(PA_i^1), g_i^2(PA_i^2), \dots, g_i^{m_i}(PA_i^{m_i})) + U_i$$

Where each V_i has bounded support ($\exists\, \inf_i, \sup_i \in \mathbb{R}$ s.t. $P(\inf_i < V_i < \sup_i) = 1$) and is taken to have m_i endogenous parents in $\boldsymbol{V} \setminus V_i$. Now, note that the assumption that all unmeasured influences, i.e. noises, can be represented by additive terms is questionable, but additive noise is commonly assumed in many causal modelling frameworks (Peters et al., 2014; Glymour, Zhang and Spirtes, 2019; Bollen et al., 2022). Therefore, we allow ourselves to assume it for the sake of our argument. In addition, we assume the weakness of the noise,

meaning that every $U_i \in \boldsymbol{U}$ is assumed to be bounded and to have a much smaller support than its corresponding endogenous variable, i.e. $|\text{supp}(U_i)| \ll |\text{supp}(V_i)|$ (we assume that all the supports are measurable, and so by the absolute value of a support we mean its Lebesgue measure). Our justifications for weakness in the discrete case extend to the continuous case: i.e. $\boldsymbol{V}$ is well-chosen, in part, when most of the mechanisms modelled via $\boldsymbol{f}$ involve causes measured by $\boldsymbol{V}$, as otherwise, we are not able to delineate the causal processes of interest in terms of quantities that we know. What's more, if we have measured enough variables in $\boldsymbol{V}$ to overcome confounding, it stands to reason as previously discussed that the remaining unmeasured sources of influence should be of limited significance, i.e. weak. Now, a function f_i is discretisable if $\forall PA_i^j \in PA_i \; \exists g_i^j : \mathbb{R} \to \mathbb{R}$ s.t. $\text{supp}(g_i^j(PA_i^j))$ is a *finite* set, that is, where g_i^j is a step function from the reals to a finite set of real numbers, such that f_i can be expressed as the composition of some function q_i with these step functions (plus noise). Note, we say that an MCM is discretisable if every $f_i \in \boldsymbol{f}$ is discretisable.

Now, granted our assumption of a discretisable MCM with weak additive noise, we argue that if each $f_i \in \boldsymbol{f}$ is randomly specified, then the probability of obtaining FVs is positive. Let $V_j \in PA_i$, s.t. $V_i := q_i(g_i^1(PA_i^1), \dots, g_i^j(V_j), \dots, g_i^{m_i}(PA_i^{m_i})) + U_i$. Now $\text{supp}\left(g_i^j(V_j)\right) = \{l_1, l_2, \dots, l_{v_i^j - 1}, l_{v_i^j}\}$, where v_i^j is the number of levels that the step function g_i^j has (that are supported). Now, each $l_t \in = \{l_1, \dots, l_{v_i^j}\}$ has a corresponding interval int_t with length d_t. Now, by the finiteness of the support, there will be a smallest interval int_{min} with length d_{min}. Next, let the noise U_j be weak enough s.t. $\left|\text{supp}(U_j)\right| < d_{min}/2$. Now, by the discretisability of f_j, and by the additivity of the noise term U_j, we have that $V_j := f_j(PA_j, U_j) = q_j(g_j^1(PA_j^1), g_j^2(PA_j^2), \dots, g_j^{m_j}(PA_j^{m_j})) + U_j$. Importantly, we may note that q_j has a range that is a finite subset of $\mathbb{R}$, as q_j's domain itself is a finite set by virtue of every endogenous parent being mapped through a step function. So, an arbitrary specification of f_j can be obtained as follows: firstly, identify the infima and suprema bounds inf_j and sup_j of V_j, which will exist by assumption. Next, let the output of q_j for a given input, be randomly specified via sampling from any probability density will full support over the interval $(\text{inf}_j, \text{sup}_j) \subset \mathbb{R}$. It readily follows that there is a positive probability of sampling a point that strictly lies at least $d_{min}/2$ away from the nearest interval boundary int_t from $\{\text{int}_1, \dots, \text{int}_{v_i^j}\}$. This can be shown by considering how any probability density on $(\text{inf}_j, \text{sup}_j)$ provides a well-defined probability for sampling a point within $d_{min}/2$ from any interval boundary, and so our value of interest is the complement of this probability. Repeating this process for every input yields a finite set of "$\left|\text{supp}(g_j^1(PA_j^1)\right| \times \dots \times \left|\text{supp}(g_j^{m_j}(PA_j^{m_j})\right|$" many probabilities, the product of which will be a positive value bounding from below the probability of sampling a f_j s.t. $\forall pa_j \in PA_j$ the term $g_i^j(V_j)$ is constant conditioned on PA_j, i.e. we have effect determinability. This in turn means that there is a positive probability that V_j and V_i will be CI w.r.t PA_j, presenting yet again a FV (due to their adjacency). Critique about why we assume full support over $(\text{inf}_j, \text{sup}_j)$ when specifiying f_j is valid, but a full discussion lies beyond the scope of this thesis. Yet in short, granted our assumptions, this approach corresponds with taking the most general view towards the form of f_j, and coheres with the principle of maximum entropy.

2.4 FVs in synthetic MCMs

For each of the settings detailed above, we generated 125,000 random MCMs across different model hyperparameters and assessed the prevalence of FVs. For our first setting characterised by discrete endogenous and exogenous variable, we considered models with 2, 4, 6, 8 and 10 endogenous nodes, where the size of the support of each endogenous node/variable was independently and randomly chosen by sampling a shifted geometric distribution with a model-wide hyperparameter for its mean set to either 3, 4, 5, 6, or 7 (shifted by adding 1 to the random variable to avoid degenerate singular support sizes). The size of the supports for the exogenous variables were determined in the same manner according to a different shifted geometric distribution with its mean also set to either 3, 4, 5, 6, or 7. Now for each of the 125 possible combinations of these hyperparameters, 1000 models were then generated with randomised DAG structures, where the probability for any edge permitted by the topological ordering of the endogenous variables was independently set to 0.3. Furthermore, the probability mass function for an exogenous variable over its support was determined by sampling a flat Dirichlet distribution with as many categories as the size of the support. Note, the usage of Dirichlet and geometric distributions follows the principle of holding minimal additional assumptions about the probabilistic structures of our models, as the geometric distribution maximises entropy over its support and a flat Dirichlet distribution gives equal weighting to every possible probability mass function over a given finite support. Now, for every given endogenous variable, its function was a randomly generated mapping from the cartesian product of the supports of the endogenous parents of the variable and its exogenous parent to its own support. Now, across our 125 hyper-parameter settings, we found that the prevalence of models with FVs (due to effect determinability) ranged from anout 0.5% for small models (i.e. 2 endogenous nodes with large supports, commonly when the means for the geometric distributions were set to 7) to 17% for large models (10 nodes) with small supports (when means were set to 3). What's more, additional analyses on synthetic data confirmed that the performance of PC's skeleton search suffers under the presence of FVs, as we found the mean accuracy of the estimated skeletons for 405 models with FVs to be about 90%, versus a mean accuracy of around 96% for 405 models with no FVs.

For our second setting with discrete endogenous variables and weak noises, we also counted the prevalence of models with FVs among 1000 randomly generated models across 125 different hyperparameter settings, with the number of endogenous variables again ranging across 2, 4, 6, 8, and 10. Now, the marginal support size of each endogenous variable was also randomly sampled according to a shifted geometric distribution; however, the mean of the distribution was kept at a fixed value of 4. Instead, we varied two new hyperparameters termed "mean weakness" and "mean simplicity" (both ranging across 0.1, 0.3, 0.5, 0.7 and 0.9). Given an endogenous variable V_i and the marginal support of an endogenous parent V_j of V_i , "mean simplicity" controlled the granularity of the mechanism for determining V_i with respect to different values for V_j by "simplifying"/filtering the support of V_j into fewer values. In particular, granted $\left|\mathrm{supp}\left(V_j\right)\right| = m$ and mean simplicity s, we sampled the shifted binomial distribution $b \sim B(m-1,1-s)+1$ (shifted to avoid values of 0), and mapped the marginal support of V_j into a reduced set of b many values. Our MCMs were then casted as causal BNs by defining for each non-root-node endogenous variable such as V_i, conditional distributions for every combination of filtered/simplified values of its parent variables. Letting $w \in (0,1)$

be the model-wide "mean simplicity" hyperparameter, conditional distributions for V_i were derived by randomly selecting c many values from $\text{supp}(V_i)$, where $c \sim B(n-1,1-w)+1$ and where $n = \text{supp}(V_i)$. A flat Dirichlet distribution with c many categories was then used to obtain the actual probabilities over these values. Note, for root node variables we only defined marginal probability distributions over their supports. Checking for the prevalence of models with FVs, we found a rate of 0% whenever model size was equal to 2 (as expected, as FVs due to effect determinability can only occur in our setting when there are at least 3 endogenous variables) to well above 90% for larger models (10 nodes) with high values for simplicity and weakness (e.g. both at 0.9). Again, FVs appeared to significantly impact PC, as an edge accuracy of about 94% was obtained for the skeletons of 360 models without FVs vs. 86% for 360 models with at least 1 or more FVs. Note, our procedure for generating DAGs (edge probabilities of 0.3) was the same as for the first setting and binomial distributions were utilised because they maximise entropy (minimising our assumptions) when given the constraints of a finite discrete support and an arbitrary mean within the range of that support.

For our third setting, we again varied model size (across the values 2, 4, 6, 8, and 10) and the hyperparameters mean weakness and mean simplicity (across 0.3, 0.4, 0.5, 0.6 and 0.7) yielding 125 hyperparameter settings, under each of which, we counted the prevalence of models with FVs out of 1000 randomly generated models. Now, granted our continuous setting, each endogenous variable was constrained, without loss of generality, to have its marginal support over a subset of $[0,1]$, and for the case of root node variables, the marginal distributions were taken to be uniform over $[0,1]$. Now, the mean simplicity s controlled how small of a finite set the values of a parent variable V_j would be mapped to, after which the mapped values would be used to determine the conditional distribution of a child variable V_i. In particular, s defined the mean of a shifted geometric distribution that was sampled to determine the number of cells that there would be in a randomly generated partition of the interval $[0,1]$, which also constituted a partition of the support of V_j. Again, geometric distributions were used due to their entropy maximising property. Now for every combination of cells from the parents of a variable V_i, a random value e from a uniform distribution over $[0,1-w]$ was chosen, where w is the model-wide weakness hyperparameter, and another random value $p \in [\frac{e}{2}, 1-\frac{e}{2}]$ was sampled, allowing us to define the conditional distribution of V_i as a uniform distribution over $[p-\frac{e}{2}, p+\frac{e}{2}]$ (in lieu of an explicit function output and an additive noise term) granted the particular combination of parent cells. Clearly, the closer the mean weakness is to 1, the narrower the conditional distribution will have to be. Now, checking for the prevalence of models with FVs, we again found a rate of 0% whenever model size was equal to 2 (which again was to be expected, as FVs due to effect determinability can only occur in this setting when there are at least 3 endogenous variables) to well above 80% for larger models (10 nodes) with high values for simplicity and weakness (e.g. both at 0.7). Again, FVs appear to significantly impact PC, as an edge accuracy of about 96% was obtained for the skeleton search step on synthetic data from 360 models without FVs vs. 91% for 360 models with (at least 1 or more) FVs. Importantly, note that the restriction to uniform distributions was for the sake of convenience, and indeed any form for the conditional distributions that respect the notion of weakness (i.e. conditional supports are smaller than marginal supports) would essentially yield the same results. Finally, for all 3 settings, we have attached visualisations from our experimental analyses in an appendix and have separately, also provided all of the code required to fully reproduce our experimental results.

Chapter 3

Conclusions

We have shown 3 settings under which FVs due to effect determinability occur with positive probabilities. With each setting being contingent on assumptions that may or may not be satisfied in a real-world context, we cannot draw any conclusions about how common FVs are in practice, as an answer to such a question will depend on how commonly our settings' assumptions our satisfied, an empirical question that may be far too difficult to answer in any generality. Nonetheless, we have argued that our assumptions are conceivable; for example, the weakness of exogenous noises fits well together with the assumption of causal sufficiency and coheres in general with the idea of having a well measured system that we can model accurately. Other assumptions such as the simplicity of a modelling function f_i with respect to a parent, or the assumption of the discretisability of $\boldsymbol{f}$ in the continuous case, also appear to us to be conceivable, and we have considered real world mechanisms, e.g. threshold mechanisms, as examples where these assumptions may be satisfied. In any case, other standard assumptions, such as acyclicity, sufficiency and the additivity of noise, may be just as, if not more, contentious and contingent than our additional assumptions. Therefore, the main takeaway from our work should be that "0 probability" arguments for FVs are very much dependent on the setting we assume we are working with, and under certain conditions, the results simply will not apply, and so, works such as Meek's (1995) should not be taken as a guarantee that we in practice can heedlessly assume the FC condition just because it has been shown that FV are extremely rare in a very general setting for discrete BNs; for granted certain additional assumptions, FVs may be more common than anticipated, as we have argued both theoretically, by deriving lower bounds on the prevalence of FVs, and experimentally by generating models according to our settings, and assessing how frequently these models present FVs. Indeed, to the best of our knowledge, our work presents the first exploration of settings under which strict FVs (e.g. as opposed to "near FVs" due to near path cancellations) are shown to occur with positive probability.

One possible line of future research pertains to how much further we can generalise out settings by weakening or generalising our sets of assumptions. We have already delved into some of the limitations of our assumptions, e.g. the variability of the threshold value in the case of simple modelling functions (with respect to an endogenous parent) and have pointed out a line of argumentation that allows us to handle such a seeming limitation. Indeed, we believe that a much more detailed and rigorous exposition of our assumptions and settings would allow us to try and deal with many addressed and unaddressed apparent limitations. In this work, we attempted to balance the level of detail presented for our results with the need to cover

numerous background topics on causality, BNs and causal discovery, without a discussion of which, it would be hard to contextualise and make sense of our results. Finally, another line of work that would, in some sense, complement or expand upon what we have done is the study of how frequent near-FVs due to the "near determinability" of effects are in our classes of discrete and discretisable MCMs (or in other similar classes of MCMs). The idea is that near violations of the FC condition will have, in practice with limited data, the same implications as true FVs for the purpose of causal discovery, in that they may induce one to assume spurious causal independences. However, near-FVs have been shown to be much more abundant, in a measure-theoretic sense, compared to strict FVs, at least in the context of FVs in linear systems due to path cancellations (Uhler et al., 2013). We expect that the same patterns will hold in our settings, with near-FVs being significantly more common, reinforcing the idea that FVs, or rather that near-FVs, may present an even greater problem in practice than assumed. In conclusion, there appear to be several avenues for future work via which we could strengthen and broaden our results, yet, as it stands, we feel confident in stating that our current work presents a small but, to the best of our knowledge, novel contribution to the study of the causal faithfulness assumption and its violability.

Bibliography

Andersen, H., 2013. When to Expect Violations of Causal Faithfulness and Why It Matters. *Philosophy of Science* [Online], 80(5), pp.672–683. Available from: https://doi.org/10.1086/673937.

Armstrong, D.M., 1999. The Open Door: Counterfactual Versus Singularist Theories of Causation. In: H. Sankey, ed. *Causation and Laws of Nature* [Online]. Australasian Studies in History and Philosophy of Science. Dordrecht: Springer Netherlands, pp.175–185. Available from: https://doi.org/10.1007/978-94-015-9229-1_16 [Accessed 28 February 2024].

Beebee, H., Hitchcock, C. and Menzies, P., 2009. *The Oxford Handbook of Causation* [Online]. Oxford University Press. Available from: https://doi.org/10.1093/oxfordhb/9780199279739.001.0001.

Bengio, Y., Deleu, T., Rahaman, N., Ke, N.R., Lachapelle, S., Bilaniuk, O., Goyal, A. and Pal, C., 2019. *A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms* [Online]. Available from: https://openreview.net/forum?id=ryxWIgBFPS [Accessed 3 October 2023].

Bollen, K.A., Fisher, Z., Lilly, A., Brehm, C., Luo, L., Martinez, A. and Ye, A., 2022. Fifty years of structural equation modeling: A history of generalization, unification, and diffusion. *Social Science Research* [Online], 107, p.102769. Available from: https://doi.org/10.1016/j.ssresearch.2022.102769.

Bongers, S., Forré, P., Peters, J. and Mooij, J.M., 2021. Foundations of Structural Causal Models with Cycles and Latent Variables. *The Annals of Statistics* [Online], 49(5). Available from: https://doi.org/10.1214/21-AOS2064 [Accessed 16 August 2023].

Briggs, R., 2012. Interventionist Counterfactuals. *Philosophical Studies* [Online], 160(1), pp.139–166. Available from: https://doi.org/10.1007/s11098-012-9908-5.

Carloni, G., Berti, A. and Colantonio, S., 2023. *The role of causality in explainable artificial intelligence* [Online]. arXiv. Available from: http://arxiv.org/abs/2309.09901 [Accessed 23 January 2024].

Chicharro, D. and Panzeri, S., 2014. Algorithms of causal inference for the analysis of effective connectivity among brain regions. *Frontiers in Neuroinformatics* [Online], 8, p.64. Available from: https://doi.org/10.3389/fninf.2014.00064.

Davidson, D., 1969. The Individuation of Events. In: N. Rescher, ed. *Essays in Honor of Carl G. Hempel*. Reidel, pp.216–34.

Dawid, A.P., 1979. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society: Series B (Methodological)* [Online], 41(1), pp.1–15. Available from: https://doi.org/10.1111/j.2517-6161.1979.tb01052.x.

Dowe, P., 2009. Causal Process Theories. In: H. Beebee, C. Hitchcock and P. Menzies, eds. *The Oxford Handbook of Causation*. Oxford University Press.

Dretske, F.I., 1977. Referring to Events. *Midwest Studies in Philosophy* [Online], 2(1), pp.90–99. Available from: https://doi.org/10.1111/j.1475-4975.1977.tb00030.x.

Engle, R.F., Hendry, D.F. and Richard, J.-F., 1983. Exogeneity. *Econometrica*, 51(2), pp.277–304.

Falcon, A., 2023. Aristotle on Causality. In: E.N. Zalta and U. Nodelman, eds. *The Stanford Encyclopedia of Philosophy* [Online]. Spring 2023. Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/archives/spr2023/entries/aristotle-causality/ [Accessed 23 January 2024].

Fisher, F.M., 1970. A Correspondence Principle for Simultaneous Equation Models. *Econometrica* [Online], 38(1), pp.73–92. Available from: https://doi.org/10.2307/1909242.

Forré, P. and Mooij, J.M., 2017. *Markov Properties for Graphical Models with Cycles and Latent Variables* [Online]. arXiv. Available from: https://doi.org/10.48550/arXiv.1710.08775 [Accessed 21 January 2024].

Frisch, M., 2022. Causation in Physics. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Spring 2022. Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/archives/spr2022/entries/causation-physics/ [Accessed 19 August 2023].

Friston, K.J., 2011. Functional and Effective Connectivity: A Review. *Brain Connectivity* [Online], 1(1), pp.13–36. Available from: https://doi.org/10.1089/brain.2011.0008.

Friston, K.J., Harrison, L. and Penny, W., 2003. Dynamic causal modelling. *NeuroImage* [Online], 19(4), pp.1273–1302. Available from: https://doi.org/10.1016/S1053-8119(03)00202-7.

Friston, K.J., Li, B., Daunizeau, J. and Stephan, K.E., 2011. Network discovery with DCM. *Neuroimage* [Online], 56(3–4), pp.1202–1221. Available from: https://doi.org/10.1016/j.neuroimage.2010.12.039.

Gallow, J.D., 2022. The Metaphysics of Causation. In: E.N. Zalta and U. Nodelman, eds. *The Stanford Encyclopedia of Philosophy* [Online]. Fall 2022. Metaphysics Research Lab, Stanford University. Available from:

https://plato.stanford.edu/archives/fall2022/entries/causation-metaphysics/ [Accessed 1 January 2024].

Geiger, D. and Pearl, J., 1990. On the Logic of Causal Models. In: R.D. Shachter, T.S. Levitt, L.N. Kanal and J.F. Lemmer, eds. *Uncertainty in Artificial Intelligence* [Online], 9. Machine Intelligence and Pattern Recognition. North-Holland, pp.3–14. Available from: https://doi.org/10.1016/B978-0-444-88650-7.50006-8.

Geiger, D., Verma, T. and Pearl, J., 1990. Identifying independence in bayesian networks. *Networks* [Online], 20(5), pp.507–534. Available from: https://doi.org/10.1002/net.3230200504.

Ghahramani, Z., 1998. Learning dynamic Bayesian networks. In: C.L. Giles and M. Gori, eds. *Adaptive Processing of Sequences and Data Structures: International Summer School on Neural Networks "E.R. Caianiello" Vietri sul Mare, Salerno, Italy September 6–13, 1997 Tutorial Lectures* [Online]. Berlin, Heidelberg: Springer, pp.168–197. Available from: https://doi.org/10.1007/BFb0053999 [Accessed 18 May 2024].

Glymour, C., Zhang, K. and Spirtes, P., 2019. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics* [Online], 10. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2019.00524 [Accessed 3 October 2023].

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y. and Schölkopf, B., 2020. *Recurrent Independent Mechanisms* [Online]. arXiv. Available from: http://arxiv.org/abs/1909.10893 [Accessed 12 October 2023].

Granger, C.W.J., 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* [Online], 37(3), pp.424–438. Available from: https://doi.org/10.2307/1912791.

Hall, N., 2004. Two Concepts of Causation. In: J. Collins, N. Hall and L. Paul, eds. *Causation and Counterfactuals*. MIT Press, pp.225–276.

Hausman, D.M. and Woodward, J., 1999. Independence, Invariance and the Causal Markov Condition. *The British Journal for the Philosophy of Science*, 50(4), pp.521–583.

Herwartz, H. and Maxand, S., 2020. Nonparametric tests for independence: a review and comparative simulation study with an application to malnutrition data in India. *Statistical Papers* [Online], 61(5), pp.2175–2201. Available from: https://doi.org/10.1007/s00362-018-1026-9.

Hitchcock, C., 2001. The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy* [Online], 98(6), pp.273–299. Available from: https://doi.org/10.2307/2678432.

Hitchcock, C., 2018. *Causal Models* [Online]. Available from: https://plato.stanford.edu/archives/fall2018/entries/causal-models/#ActuCaus [Accessed 19

August 2023].

Hitchcock, C., 2023. Causal Models. In: E.N. Zalta and U. Nodelman, eds. *The Stanford Encyclopedia of Philosophy* [Online]. Spring 2023. Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/archives/spr2023/entries/causal-models/ [Accessed 2 January 2024].

Hitchcock, C. and Rédei, M., 2021. Reichenbach's Common Cause Principle. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Summer 2021. Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/archives/sum2021/entries/physics-Rpcc/ [Accessed 14 March 2024].

Holland, P.W., 1986. Statistics and Causal Inference. *Journal of the American Statistical Association* [Online], 81(396), pp.945–960. Available from: https://doi.org/10.1080/01621459.1986.10478354.

Janzing, D., Peters, J., Mooij, J. and Schölkopf, B., 2009. Identifying confounders using additive noise models. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* [Online], Arlington, Virginia, USA. UAI '09. Arlington, Virginia, USA: AUAI Press, pp.249–257. Available from: [Accessed 22 December 2023].

Kim, J., 1976. Events as Property Exemplifications. In: M. Brand and D. Walton, eds. *Action Theory*. D. Reidel, pp.310–326.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N. and Leimer, H.-G., 1990. Independence properties of directed markov fields. *Networks* [Online], 20(5), pp.491–505. Available from: https://doi.org/10.1002/net.3230200503.

Lemeire, J., Steenhaut, K. and Maes, S., 2012. *Causal Inference on Data Containing Deterministic Relations.*

Lewis, D., 1973a. Causation. *Journal of Philosophy* [Online], 70(17), pp.556–567. Available from: https://doi.org/10.2307/2025310.

Lewis, D.K., 1973b. *Counterfactuals*. Malden, Mass.: Blackwell.

Li, L., Ghothani, H.A., Huang, B., Chen, G., Zhang, J., Harel, S. and Zhang, K., 2023. *On Causal Discovery in the Presence of Deterministic Relations* [Online]. Available from: https://openreview.net/forum?id=jE6VXUhxq9&referrer=%5Bthe%20profile%20of%20Jiji%20Zhang%5D(%2Fprofile%3Fid%3D~Jiji_Zhang1)# [Accessed 23 April 2024].

Lombard, M. and Gärdenfors, P., 2017. Tracking the evolution of causal cognition in humans. *Journal of Anthropological Sciences* [Online], (95), pp.219–234. Available from: https://doi.org/10.4436/JASS.95006.

Mabrouk, A., Gonzales, C., Jabet-Chevalier, K. and Chojnacki, E., 2014. An Efficient

Bayesian Network Structure Learning Algorithm in the Presence of Deterministic Relations. *ECAI 2014* [Online]. IOS Press, pp.567–572. Available from: https://doi.org/10.3233/978-1-61499-419-0-567 [Accessed 10 April 2024].

MacCallum, R.C. and Austin, J.T., 2000. Applications of Structural Equation Modeling in Psychological Research. *Annual Review of Psychology* [Online], 51(Volume 51, 2000), pp.201–226. Available from: https://doi.org/10.1146/annurev.psych.51.1.201.

Mackie, J.L., 1980. *The Cement of the Universe: A Study of Causation* [Online]. Oxford University Press. Available from: https://doi.org/10.1093/0198246420.001.0001 [Accessed 3 May 2024].

Meek, C., 1995. Strong completeness and faithfulness in Bayesian networks. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* [Online], San Francisco, CA, USA. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp.411–418. Available from: [Accessed 11 May 2024].

Menzies, P. and Beebee, H., 2020. Counterfactual Theories of Causation. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Winter 2020. Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/archives/win2020/entries/causation-counterfactual/ [Accessed 23 January 2024].

Mihajlovic, V. and Petkovic, M., 2001. *Dynamic Bayesian Networks: A State of the Art* [Online]. University of Twente. Available from: https://research.utwente.nl/en/publications/dynamic-bayesian-networks-a-state-of-the-art [Accessed 18 May 2024].

Mooij, J.M., Janzing, D. and Schölkopf, B., 2013. *From Ordinary Differential Equations to Structural Causal Models: the deterministic case* [Online]. arXiv. Available from: http://arxiv.org/abs/1304.7920 [Accessed 3 October 2023].

Okamoto, M., 1973. Distinctness of the Eigenvalues of a Quadratic form in a Multivariate Sample. *The Annals of Statistics*, 1(4), pp.763–765.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M. and Schölkopf, B., 2018. *Learning Independent Causal Mechanisms* [Online]. arXiv. Available from: https://doi.org/10.48550/arXiv.1712.00961 [Accessed 5 March 2024].

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J., 2009. *Causality* [Online]. 2nd ed. Cambridge: Cambridge University Press. Available from: https://doi.org/10.1017/CBO9780511803161 [Accessed 29 February 2024].

Pearl, J. and Mackenzie, D., 2018. *The Book of Why: The New Science of Cause and Effect*. 1st ed. USA: Basic Books, Inc.

Peters, J., 2014. *On the Intersection Property of Conditional Independence and its Application to Causal Discovery* [Online]. arXiv. Available from: http://arxiv.org/abs/1403.0408 [Accessed 23 April 2024].

Peters, J., Bühlmann, P. and Meinshausen, N., 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* [Online], 78(5), pp.947–1012. Available from: https://doi.org/10.1111/rssb.12167.

Peters, J., Mooij, J.M., Janzing, D. and Schölkopf, B., 2014. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research*, 15(58), pp.2009–2053.

Quine, W.V., 1985. Events and Reification. In: E. Lepore and B. McLaughlin, eds. *Actions and Events: Perspectives on the Philosophy of Davidson*. Blackwell, pp.162–71.

Rehder, B., 2017. Reasoning With Causal Cycles. *Cognitive Science* [Online], 41(S5), pp.944–1002. Available from: https://doi.org/10.1111/cogs.12447.

Reichenbach, H., 1956. *The Direction of Time*. Ed. by M. Reichenbach. Mineola, N.Y.: Dover Publications.

Russell, B., 1912. On the Notion of Cause. *Proceedings of the Aristotelian Society*, 13, pp.1–26.

Sadeghi, K., 2017. Faithfulness of probability distributions and graphs. *The Journal of Machine Learning Research*, 18(1), pp.5429–5457.

Sanchez-Romero, R., Ramsey, J.D., Zhang, K., Glymour, M.R.K., Huang, B. and Glymour, C., 2019. Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Network Neuroscience* [Online], 3(2), pp.274–306. Available from: https://doi.org/10.1162/netn_a_00061.

Schaffer, J., 2005. Contrastive Causation. *Philosophical Review* [Online], 114(3), pp.327–358. Available from: https://doi.org/10.1215/00318108-114-3-327.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A. and Bengio, Y., 2021. *Towards Causal Representation Learning* [Online]. arXiv. Available from: http://arxiv.org/abs/2102.11107 [Accessed 12 October 2023].

Shajarisales, N., Janzing, D., Schoelkopf, B. and Besserve, M., 2015. Telling cause from effect in deterministic linear dynamical systems. In: F. Bach and D. Blei, eds. *Proceedings of the 32nd International Conference on Machine Learning* [Online], Lille, France, 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp.285–294. Available from: https://proceedings.mlr.press/v37/shajarisales15.html.

Shojaie, A. and Fox, E.B., 2022. Granger Causality: A Review and Recent Advances.

Annual Review of Statistics and Its Application [Online], 9(1), pp.289–319. Available from: https://doi.org/10.1146/annurev-statistics-040120-010930.

Spirtes, P., 1993. *Directed cyclic graphs, conditional independence, and non-recursive linear structural equation models* [Online]. Carnegie Mellon University. Available from: https://doi.org/10.1184/R1/6491399.v1 [Accessed 4 October 2023].

Spirtes, P. and Glymour, C., 1991. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review* [Online], 9(1), pp.62–72. Available from: https://doi.org/10.1177/089443939100900106.

Spirtes, P. and Zhang, K., 2016. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics* [Online], 3(1), p.3. Available from: https://doi.org/10.1186/s40535-016-0018-x.

Spirtes, Peter, Glymour, Clark, and Scheines, Richard, 2001. *Causation, Prediction, and Search, Second Edition* [Online]. 2nd ed. The MIT Press. Available from: https://mitpress.mit.edu/9780262527927/causation-prediction-and-search/ [Accessed 11 October 2023].

Spohn, W., 1980. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic* [Online], 9(1), pp.73–99. Available from: https://doi.org/10.1007/BF00258078.

Steel, D., 2006. Homogeneity, selection, and the faithfulness condition. *Minds and Machines* [Online], 16(3), pp.303–317. Available from: https://doi.org/10.1007/s11023-006-9032-4.

Tigas, P., Annadani, Y., Jesson, A., Schölkopf, B., Gal, Y. and Bauer, S., 2022. Interventions, Where and How? Experimental Design for Causal Models at Scale. *Advances in Neural Information Processing Systems*, 35, pp.24130–24143.

Tiles, J.E., 1993. Experiment as Intervention. *The British Journal for the Philosophy of Science* [Online], 44(3), pp.463–475. Available from: https://doi.org/10.1093/bjps/44.3.463.

Toth, C., Lorch, L., Knoll, C., Krause, A., Pernkopf, F., Peharz, R. and von Kügelgen, J., 2022. Active Bayesian Causal Inference. *Advances in Neural Information Processing Systems*, 35, pp.16261–16275.

Uhler, C., Raskutti, G., Bühlmann, P. and Yu, B., 2013. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics* [Online], 41(2). Available from: https://doi.org/10.1214/12-AOS1080 [Accessed 27 March 2024].

Verma, T. and Pearl, J., 1990. Causal Networks: Semantics and Expressiveness*. In: R.D. Shachter, T.S. Levitt, L.N. Kanal and J.F. Lemmer, eds. *Machine Intelligence and Pattern Recognition* [Online], 9. Uncertainty in Artificial Intelligence. North-Holland, pp.69–76. Available from: https://doi.org/10.1016/B978-0-444-88650-7.50011-1 [Accessed 9 April 2024].

Verma, T. and Pearl, J., 2022. Equivalence and Synthesis of Causal Models. *Probabilistic and Causal Inference: The Works of Judea Pearl* [Online], 36. 1st ed. New York, NY, USA: Association for Computing Machinery, pp.221–236. Available from: https://doi.org/10.1145/3501714.3501732 [Accessed 26 March 2024].

Waldmann, M.R., 2017. *The Oxford Handbook of Causal Reasoning* [Online]. Oxford University Press. Available from: https://doi.org/10.1093/oxfordhb/9780199399550.001.0001.

Weinberger, N., 2018. Faithfulness, Coordination and Causal Coincidences. *Erkenntnis* [Online], 83(2), pp.113–133. Available from: https://doi.org/10.1007/s10670-017-9882-6.

Weinberger, N., 2023. *Homeostasis and the Faithless Foundations of Causal Inference* [Online]. PhilSci. Available from: https://philsci-archive.pitt.edu/22261/.

Weinberger, N., n.d. *Homeostasis and the Faithless Foundations of Causal Inference*.

Woodward, J., 1989. The Causal Mechanical Model of Explanation. *Minnesota Studies in the Philosophy of Science*, 13, pp.359–83.

Woodward, J., 1995. Causation and Explanation in Econometrics. In: D. Little, ed. *On the Reliability of Economic Models: Essays in the Philosophy of Economics* [Online]. Dordrecht: Springer Netherlands, pp.9–61. Available from: https://doi.org/10.1007/978-94-011-0643-6_2.

Woodward, J. and Ross, L., 2021. Scientific Explanation. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Summer 2021. Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/ [Accessed 23 January 2024].

Yang, Y., Nafea, M., Ghassami, A. and Kiyavash, N., 2022. *Causal Discovery in Linear Structural Causal Models with Deterministic Relations* [Online]. arXiv. Available from: http://arxiv.org/abs/2111.00341 [Accessed 5 March 2024].

Zanga, A. and Stella, F., 2022. A Survey on Causal Discovery: Theory and Practice. *International Journal of Approximate Reasoning* [Online], 151, pp.101–129. Available from: https://doi.org/10.1016/j.ijar.2022.09.004.

Zeng, Y., Hao, Z., Cai, R., Xie, F., Huang, L. and Shimizu, S., 2023. Nonlinear Causal Discovery for High-Dimensional Deterministic Data. *IEEE Transactions on Neural Networks and Learning Systems* [Online], 34(5), pp.2234–2245. Available from: https://doi.org/10.1109/TNNLS.2021.3106111.

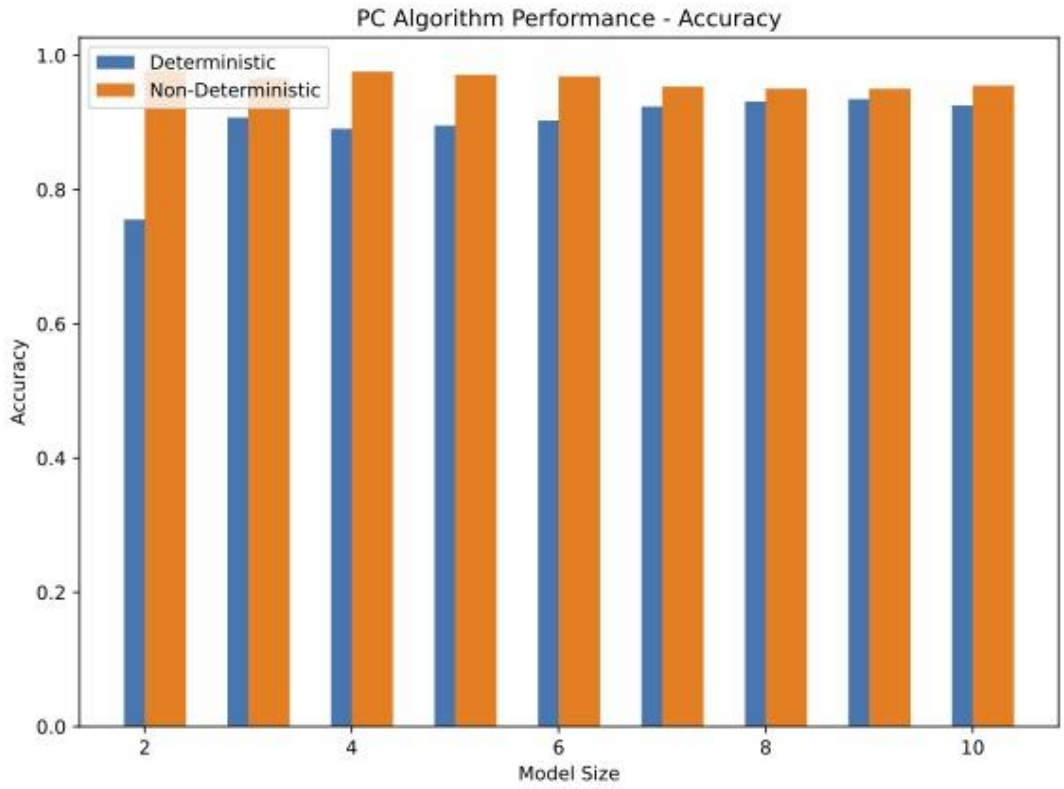
Zhang, J., Cammarata, L., Squires, C., Sapsis, T.P. and Uhler, C., 2023. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence* [Online], pp.1–10. Available from: https://doi.org/10.1038/s42256-023-00719-0.

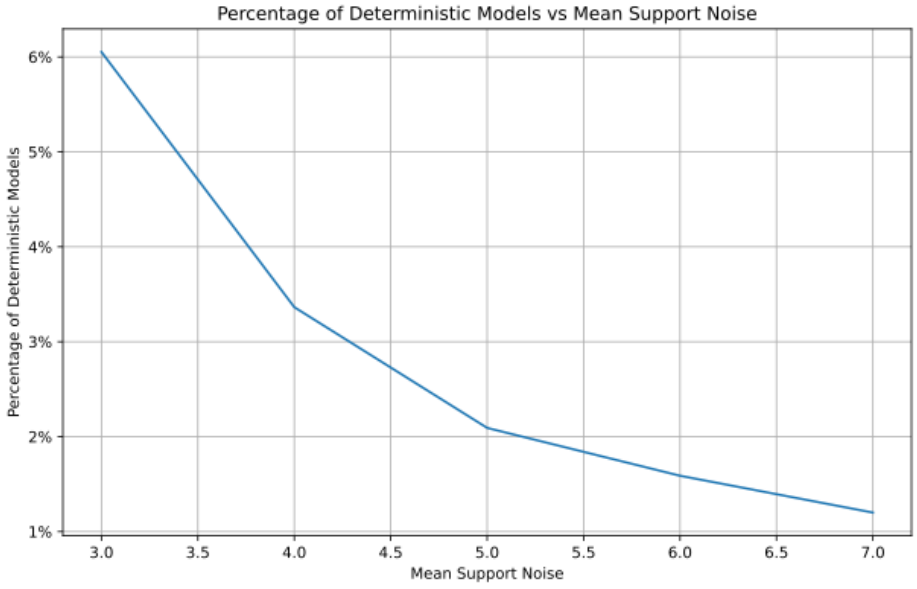
Zhang, K., Peters, J., Janzing, D. and Schoelkopf, B., 2012. *Kernel-based Conditional Independence Test and Application in Causal Discovery* [Online]. arXiv. Available from: https://doi.org/10.48550/arXiv.1202.3775 [Accessed 19 May 2024].

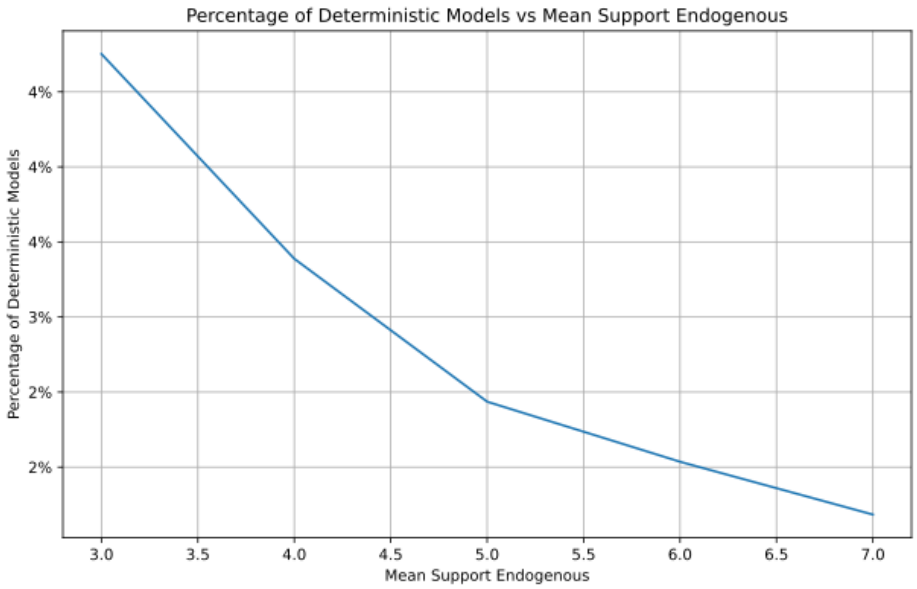
Appendix

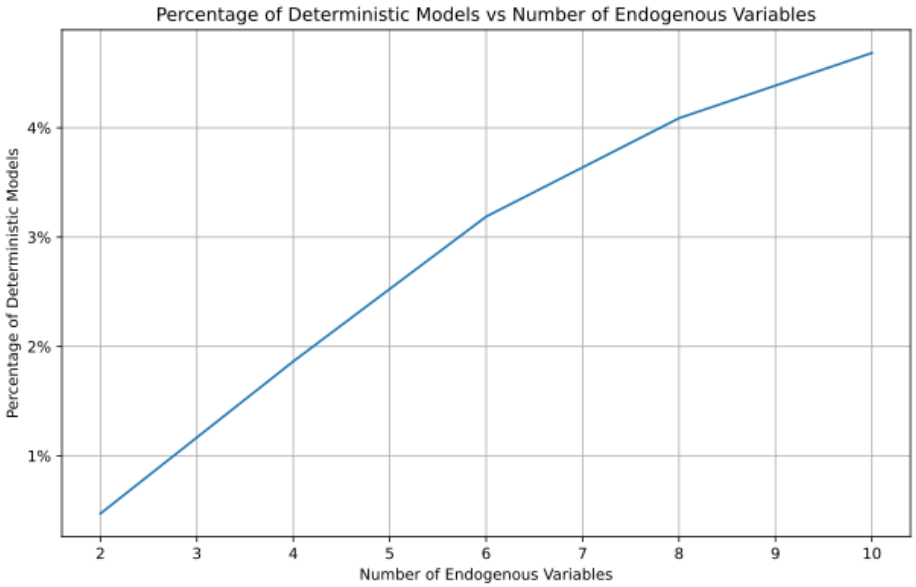
Experimental results visualisations

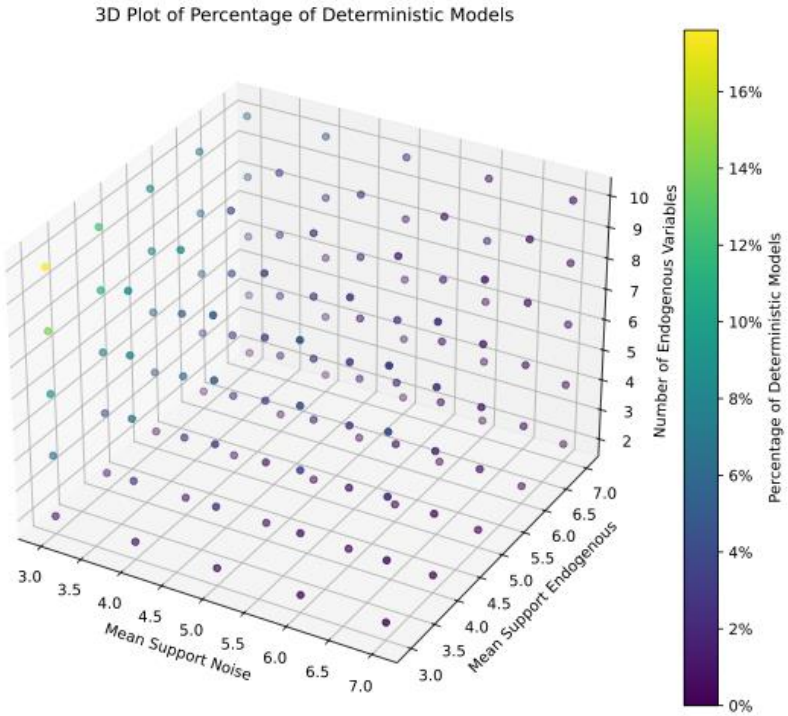
Discrete MCMs with discrete noise



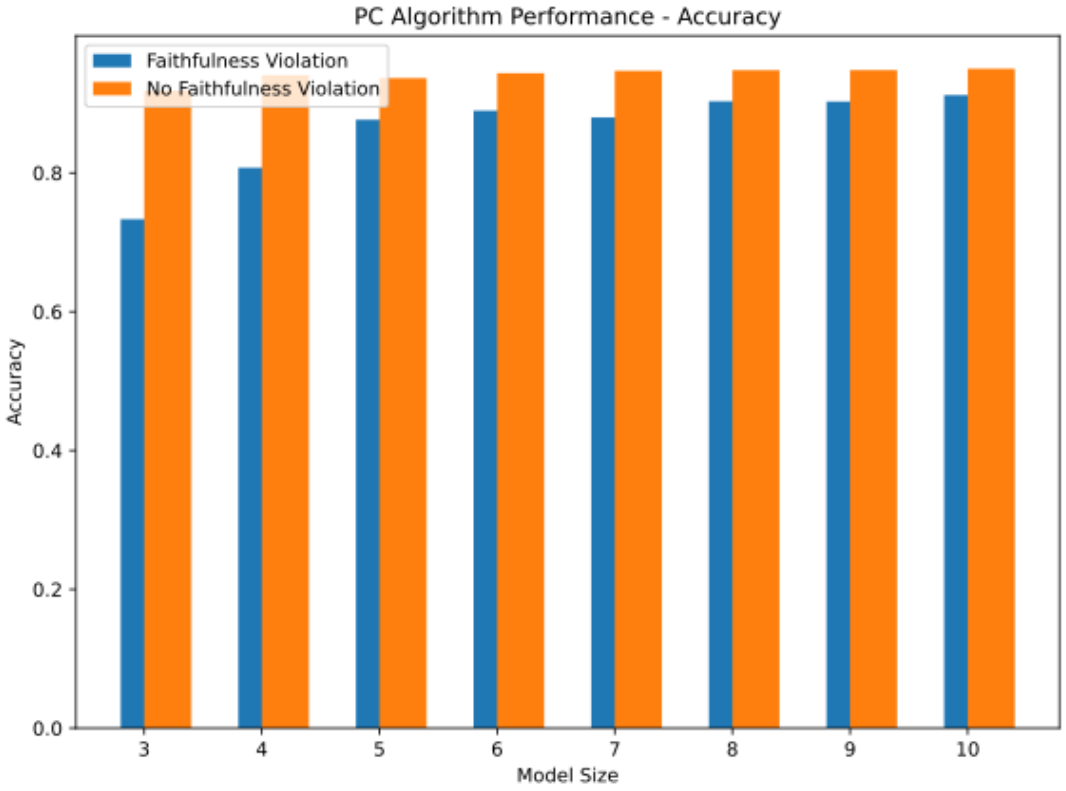

Percentage of Deterministic Models vs Mean Support Noise
Percentage of Deterministic Models
6%
5%
4%
3%
2%
1%
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
Mean Support Noise

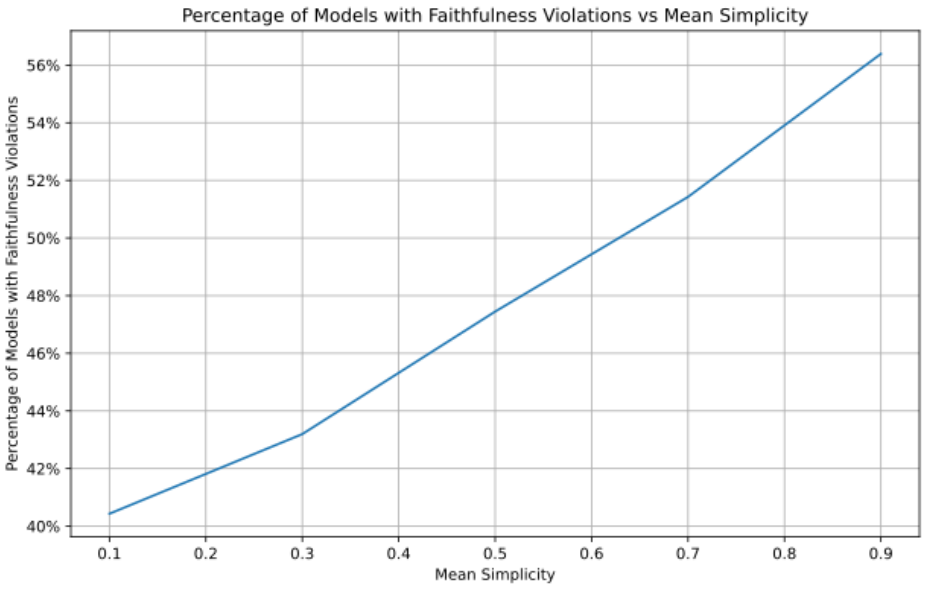

Percentage of Deterministic Models vs Mean Support Endogenous
Percentage of Deterministic Models
4%
4%
4%
3%
2%
2%
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
Mean Support Endogenous


Percentage of Deterministic Models vs Number of Endogenous Variables
Percentage of Deterministic Models
4%
3%
2%
1%
2
3
4
5
6
7
8
9
10
Number of Endogenous Variables

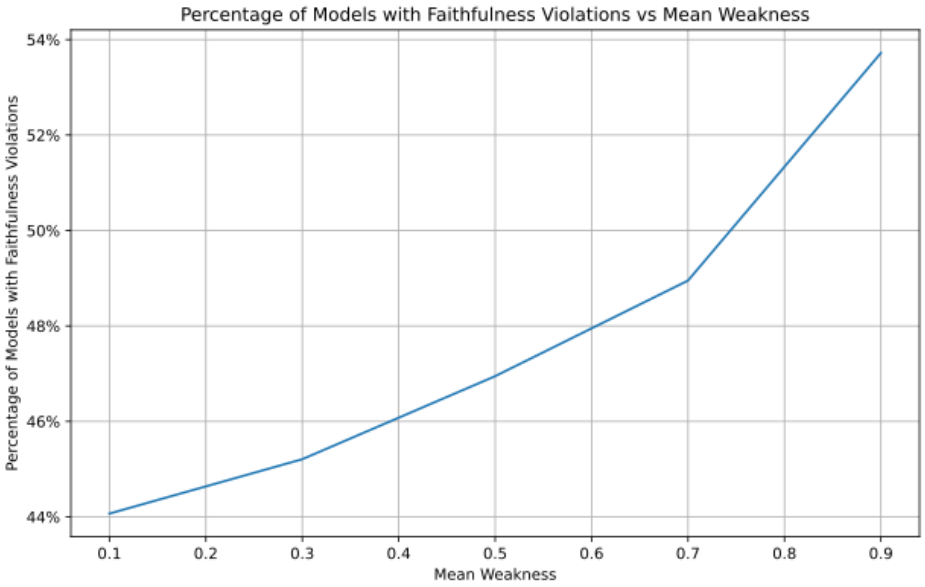

3D Plot of Percentage of Deterministic Models
Mean Support Noise
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
Mean Support Endogenous
Number of Endogenous Variables
Percentage of Deterministic Models
16%
14%
12%
10%
8%
6%
4%
2%
0%

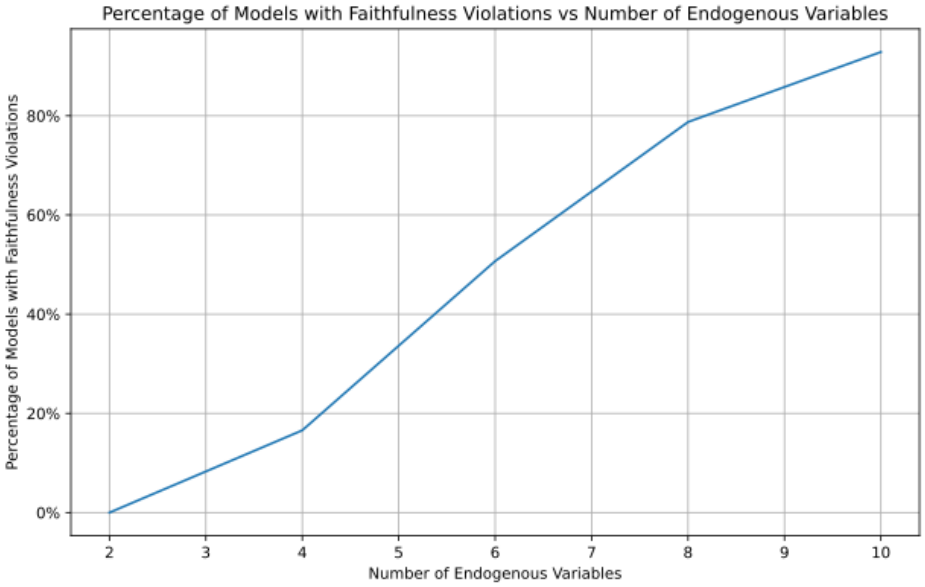
Discrete MCMs with weak noise

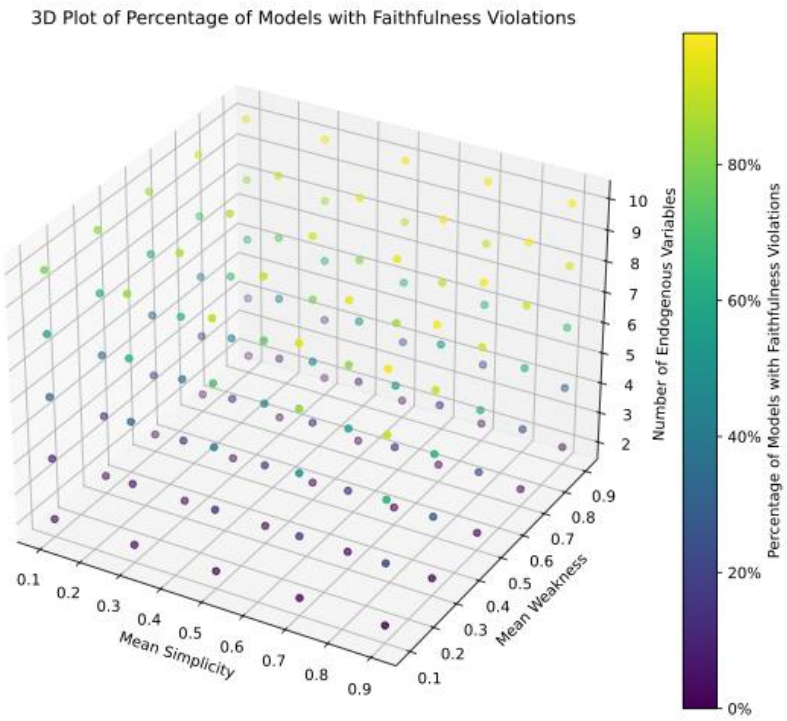




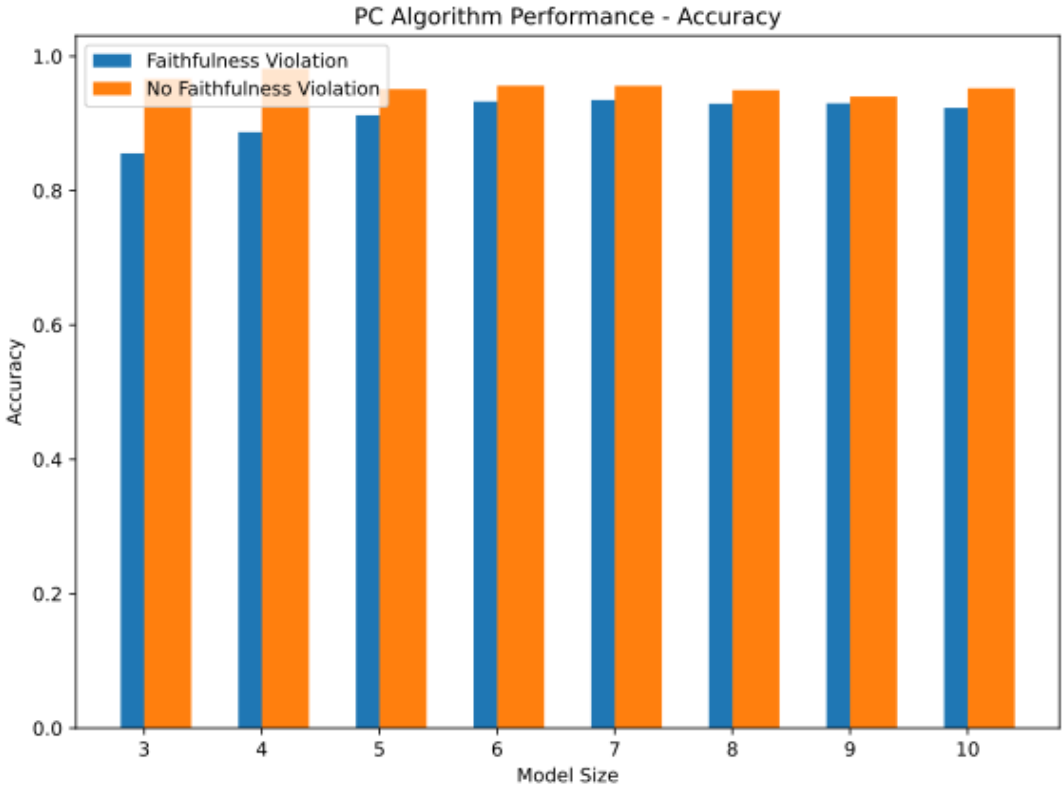
Faithfulness violations due to the determinability of effects in certain classes of discrete and discretisable Markovian causal models

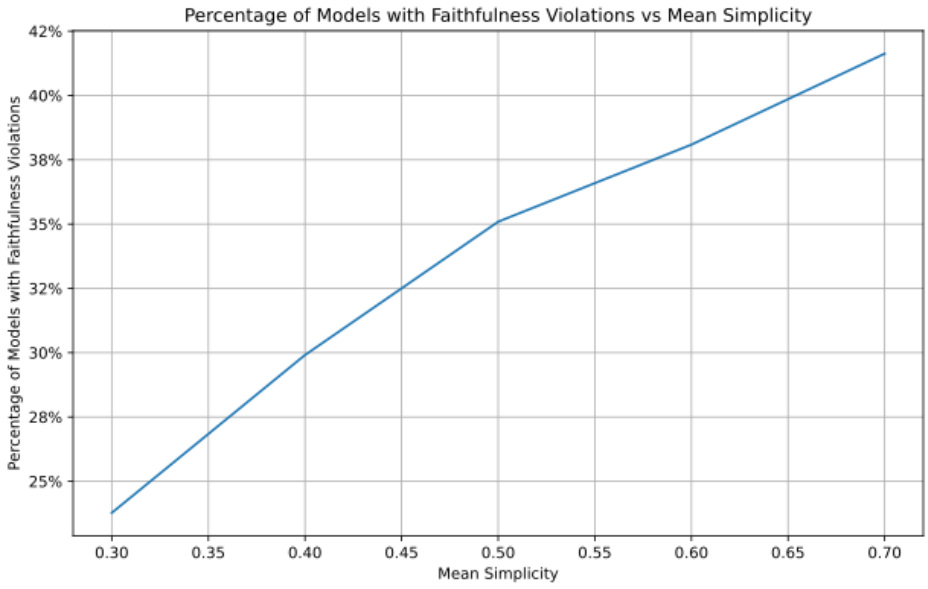


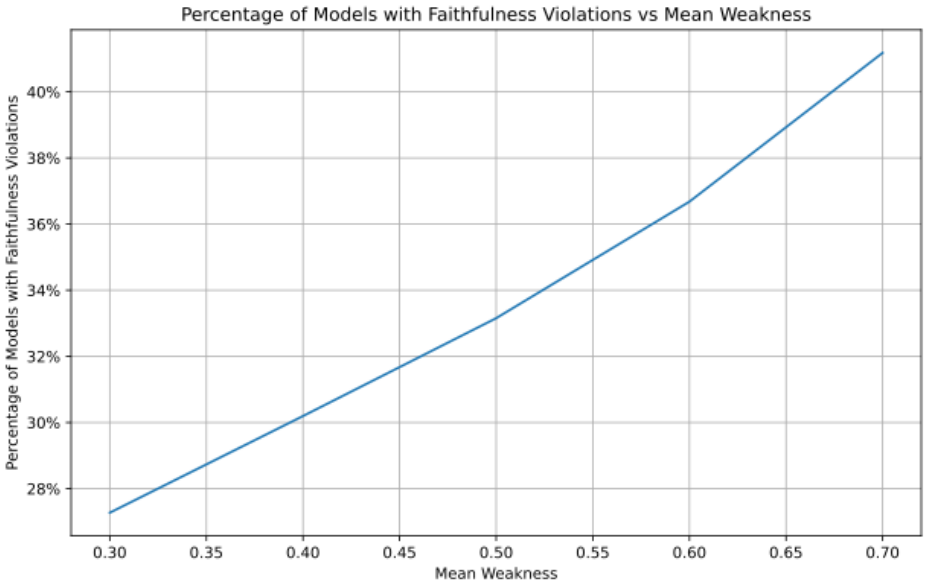


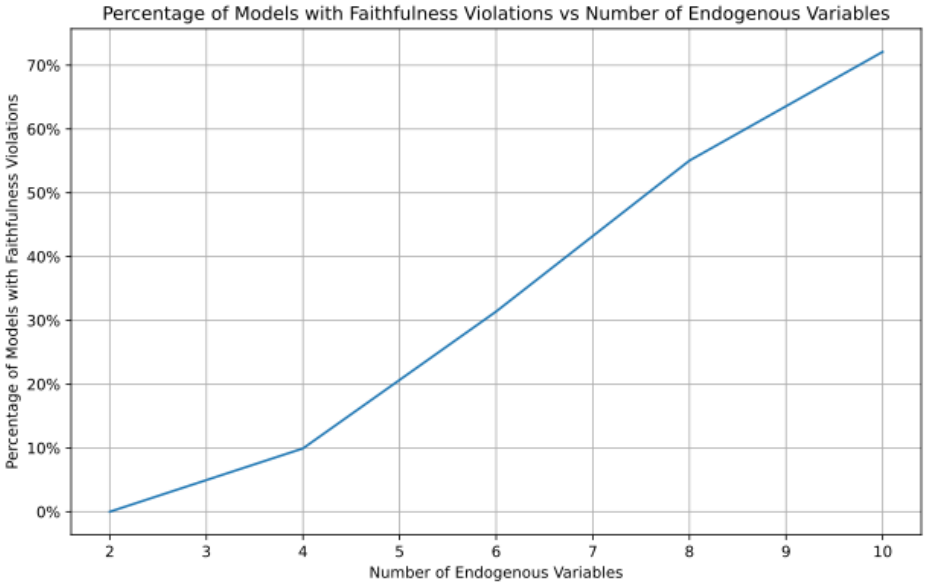

3D Plot of Percentage of Models with Faithfulness Violations
Mean Simplicity
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Mean Weakness
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Number of Endogenous Variables
2
3
4
5
6
7
8
9
10
Percentage of Models with Faithfulness Violations
0%
20%
40%
60%
80%

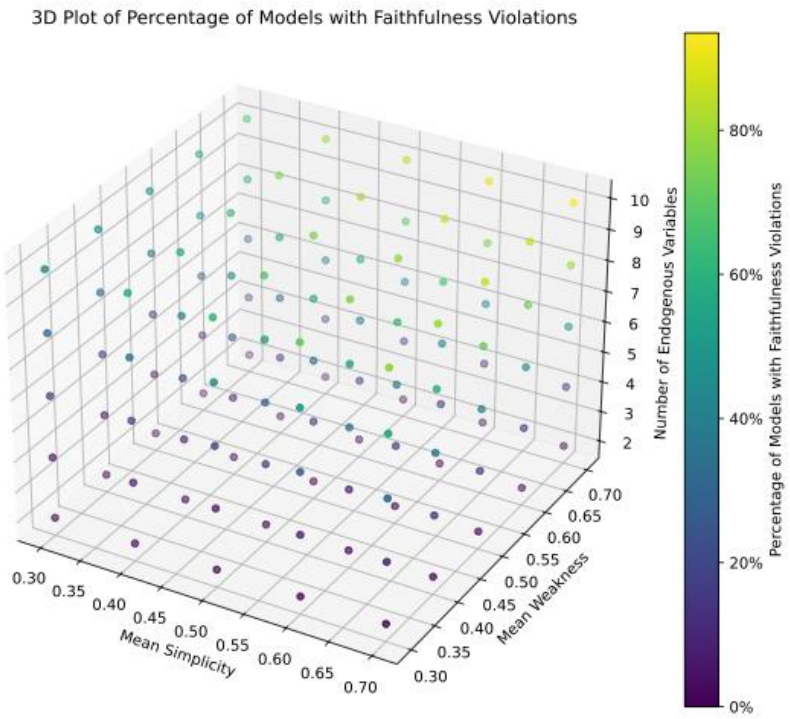
Discretisable MCMs with weak additive noise






Percentage of Models with Faithfulness Violations vs Mean Weakness
Percentage of Models with Faithfulness Violations
40%
38%
36%
34%
32%
30%
28%
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Mean Weakness


Percentage of Models with Faithfulness Violations vs Number of Endogenous Variables
Percentage of Models with Faithfulness Violations
70%
60%
50%
40%
30%
20%
10%
0%
2
3
4
5
6
7
8
9
10
Number of Endogenous Variables


3D Plot of Percentage of Models with Faithfulness Violations
Mean Simplicity
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Mean Weakness
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Number of Endogenous Variables
2
3
4
5
6
7
8
9
10
Percentage of Models with Faithfulness Violations
0%
20%
40%
60%
80%