Brit. J. Phil. Sci., **0** (2020), 1–XX
doi: XYZ

The PC Algorithm and the Inference to Constitution

Lorenzo Casini and Michael Baumgartner

Abstract

Gebharter ([2017b]) has proposed to use one of the best known Bayesian network (BN) causal discovery algorithms, PC, to identify the constitutive dependencies underwriting mechanistic explanations. His proposal assumes that mechanistic constitution behaves like deterministic direct causation, such that PC is directly applicable to mixed variable sets featuring both causal and constitutive dependencies. Gebharter claims that such mixed sets, under certain restrictions, comply with PC's background assumptions. The aim of this paper is to show that Gebharter's proposal incurs severe problems, ultimately rooted in the widespread non-compliance of mechanistic systems with PC's assumptions. This casts severe doubts on the attempt to implicitly define constitution as a form of deterministic direct causation complying with PC's assumptions.

1 Introduction

The mechanistic account of scientific explanation (Machamer, Darden, and Craver [2000]; Bechtel and Abrahamsen [2005]; Glennan [2002]) holds that the explanandum, a higher-level phenomenon, is explained by the lower-level mechanism responsible for it. In a popular characterization,

> [a] mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena. (Bechtel and Abrahamsen [2005], p. 423)

To give a simple but paradigmatic example, which shall serve as our guiding example throughout the paper, the phenomenon of amplification in a two-stage amplifier is caused by a signal (e.g., current, voltage, power) received from an input source, and causes effects such as signal distortion in an output device (e.g., a loudspeaker). The phenomenon is explained by the augmentation of the signal by the amplifier's two transistors arranged in series (see Wimsatt [2007], ch. 12).

More generally, a mechanism is embedded in a causal context, where causal background conditions are operative relative to which certain parts of the system are responsible for the phenomenon. The relevant kind of responsibility is constitutive rather than causal. The system's parts that mechanistically explain the phenomenon are the 'component' (cf. quote), or constituent, parts. While causation has been at the centre of philosophical theorizing for centuries, the notion of constitution, or constitutive relevance, has only recently begun to attract philosophical attention. In particular, it is still unclear what discovery method(s) could systematize data-based inference to constitution.

The problem of constitutive discovery is the following: given a set of spatiotemporal parts of an explanandum phenomenon, which of these parts are explanatorily relevant, that is, constituents of the phenomenon? Importantly, clarity on parthood relations (i.e., spatiotemporal overlap) between macro and micro entities is customarily assumed by all proposed solutions of this problem (Craver [2007]; Harbecke [2010]; Couch [2011]; Gebharter [2017b]; Baumgartner and Casini [2017]). Parthood itself is only necessary but not sufficient for constitution and, hence, must be complemented by additional criteria in order to identify constituents. Recently, Gebharter ([2017b]) has proposed to bring to bear PC—one of the best known causal discovery algorithms from the Bayesian network (BN) framework (Spirtes, Glymour, and Scheines [2000]; Pearl [2009])—on the task of identifying the constituents among a phenomenon's parts. To model and discover causation, PC identifies conditional independence constraints with statistical tests and, assuming that the analysed system satisfies certain BN axioms, causally connects variables not found to be independent. Gebharter claims that, despite fundamental differences between causation and constitution, constitutive relations comply with the BN axioms PC assumes for causation, such that constitution

can be methodologically treated as a form of (deterministic) direct causation. He concludes that PC may, together with parthood information, concurrently be applied to both causal and constitutive discovery in variable sets featuring both causally and constitutively related variables.

This paper critically reviews Gebharter's proposal. In a nutshell, we take issue with Gebharter's contention that constitution can be methodologically treated as a form of causation. As violations of BN axioms can be argued to be rare in variable sets exclusively featuring causal relations, which are assumed to be non-deterministic in the BN framework, these axioms are justifiably assumable for causal contexts. But constitutive relations generate deterministic dependencies, in the presence of which violations of BN axioms are no longer rare but commonplace, thus undermining their justifiable assumability and, a fortiori, the reliability of PC's output. On the basis of two benchmarking experiments, we show that even under discovery circumstances that—apart from the presence of determinism—are maximally favourable to the performance of PC, the algorithm's capacity to correctly recover constitutive relations is not high enough to counterbalance the severe risk of false positives.

The paper is organized as follows. Section 2 briefly introduces the BN framework. Section 3 reviews Gebharter's approach for PC-based constitutive discovery. Sections 4 and 5 question the justifiable assumability of BN axioms in the context of Gebharter's proposal. Section 6 considers the prospects of doing constitutive inference with a version of PC that is explicitly designed for variable sets featuring deterministic dependencies, and finds them dim. Finally, Section 7 discusses our benchmarking experiments—for which we provide a detailed replication script in the paper's supplementary material.

2 Preliminaries

We begin by introducing the BN axioms assumed by PC, as well as a notational convention on the variables of BNs representing mechanistic systems.

Traditionally, the BN formalism uses generic random variables to represent types (or degrees) of properties or behaviours independently of the entities instantiating them. Here, however, we shall follow the mechanistic literature in taking the variables as denoting the behaviours exhibited by specific entities (such as a system and its constituents), and consequently adopt the following notational convention. Calligraphic fonts are used for specific random variables $\mathcal{A}(\mathrm{S})$ and $\mathcal{B}(\mathrm{P}_1)$ (Spohn [2006]), by which we denote the behaviour $\mathcal{A}$ of a specific system S and the behaviour $\mathcal{B}$ of one specific part P_1 of S. As we are only concerned with specific variables, we will leave the entity-relativity of our variables implicit and just write '$\mathcal{A}$', '$\mathcal{B}$', *et cetera*, for the behaviour types '$\mathcal{A}(\mathrm{S})$', '$\mathcal{B}(\mathrm{P}_1)$', *et cetera*.

A BN is a triple $\langle \mathbf{V}, \mathbf{E}, \Pr \rangle$ composed of a finite set of variables $\mathbf{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_n\}$, each taking finitely many possible values, of a set of edges $\mathbf{E}$ over the variables in $\mathbf{V}$, such that variables and edges $\langle \mathbf{V}, \mathbf{E} \rangle$ form a directed acyclic graph (DAG), and of a probability dis-

tribution Pr, such that the probability of each variable $\mathcal{V}_i$ in the DAG obeys the Markov Condition (MC):

(MC) For any $\mathcal{V}_i \in \mathbf{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_n\}$, $\mathcal{V}_i \perp\!\!\!\perp \mathbf{Non}_i \mid \mathbf{Par}_i$,

where $\mathbf{Par}_i$ denotes the set of parents of $\mathcal{V}_i$, and $\mathbf{Non}_i$ denotes the set of non-descendants of $\mathcal{V}_i$. In words, each variable is probabilistically independent of its non-descendants, conditional on its parents, or equivalently, its parents screen it off from its non-descendants.

In a causally interpreted BN, the edges stand for direct causal relations, $\mathbf{Par}_i$ denotes the set of $\mathcal{V}_i$'s direct causes, $\mathbf{Non}_i$ the set of $\mathcal{V}_i$'s non-effects in the true causal structure regulating the behaviour of the variables in $\mathbf{V}$, and MC is called Causal Markov Condition (CMC) (Spirtes, Glymour, and Scheines [2000], §3.4.1, §3.5.1).

In addition to CMC, the PC algorithm assumes the Causal Faithfulness Condition (CFC) (Zhang and Spirtes [2008], p. 247):

(CFC) $\langle \mathbf{V}, \mathbf{E}, \Pr \rangle$ is such that every conditional independence relation true in Pr is entailed by CMC applied to the true DAG $\langle \mathbf{V}, \mathbf{E} \rangle$.

CFC guarantees that there is no causal dependence without a probabilistic dependence, in other words, that all probabilistic independencies in the graph are due to the absence of causal dependencies.

CMC and CFC are provably satisfied or only rarely violated in many well-known discovery contexts, guaranteeing that PC is reliably applicable to 'oracle' (true) information on conditional dependencies and independencies in those contexts.[1] On the one hand, the following is a set of conditions that are jointly sufficient (albeit not all of them are necessary) for CMC to be provably satisfied: (i) the functional relations in the data-generating structure are linear, (ii) the exogenous variables and error terms are independently distributed, (iii) all non-deterministic dependencies in the data (i.e., dependencies not producing conditional probabilities equal to 1) are due to noise and not to some fundamentally indeterministic process, meaning that all non-deterministic dependencies are so-called pseudoindeterministic, and (iv) the variable set is causally sufficient (Spirtes, Glymour, and Scheines [2000], p. 35). (Causal) Sufficiency is formulated as follows (Zhang [2006], p. 8):

(Sufficiency) For every pair of variables in $\mathbf{V}$, every common direct cause of them is also in $\mathbf{V}$ or has the same value for all units in the population.[2]

Sufficiency is necessary for the satisfaction of CMC: if Sufficiency is violated, there may be varying latent common causes, which, in turn, may induce probabilistic dependencies

[1] For a description of the algorithmic steps of PC, see Spirtes, Glymour, and Scheines ([2000], pp. 84–5).

[2] The notion of a direct common cause, in turn, is spelled out as follows: for any $\mathcal{V}_1$ and $\mathcal{V}_2$ ($\mathcal{V}_1 \neq \mathcal{V}_2$) in $\mathbf{V}$, $\mathcal{V}_3$ is a direct common cause of $\mathcal{V}_1$ and $\mathcal{V}_2$ if and only if $\mathcal{V}_3$ is a direct cause of $\mathcal{V}_1$ and a direct cause of $\mathcal{V}_2$ relative to $\mathbf{V} \cup \{\mathcal{V}_3\}$.

between variables in **V** that are spurious, that is, not due to causation and, hence, violate CMC.

On the other hand, a sufficient (but not necessary) condition for CFC to hold is that (i) and (ii) hold, and (v) the data contain no deterministic but only pseudoindeterministic dependencies. In that case, violations of CFC have Lebesgue measure 0, which entails that they can only be produced under very strong assumptions (Spirtes, Glymour, and Scheines [2000], p. 42). This, in turn, is typically taken as a reason to expect them to be very rare.

At the same time, there are well-known contexts in which BN axioms are frequently violated and, hence, not justifiably assumable. One such context, relevant for the remainder of this paper, involves deterministic dependencies in the data (which generate conditional probabilities equal to 1). In the presence of determinism, violations of CFC are commonplace (Spirtes, Glymour, and Scheines [2000], §3.8; Glymour [2007], p. 236). To illustrate, whenever the dependencies along a path $\mathcal{V}_1 \longrightarrow \mathcal{V}_2 \longrightarrow \mathcal{V}_3$ are deterministic, that is, whenever $\mathcal{V}_1$ determines $\mathcal{V}_2$, which determines $\mathcal{V}_3$, it holds that $\Pr(\mathcal{V}_3 \,|\, \mathcal{V}_1 \wedge \mathcal{V}_2) = \Pr(\mathcal{V}_3 \,|\, \mathcal{V}_1) = 1$, namely the indirect cause $\mathcal{V}_1$ screens off $\mathcal{V}_3$ from its direct cause $\mathcal{V}_2$. These screening-off relations, however, are not entailed by CMC, and hence violate CFC. That is, every deterministic chain violates CFC. The systematicity of CFC violations under determinism entails that PC is not justifiably applicable to deterministic data.

3 Gebharter's Proposal

While PC is one of the most frequently discussed causal discovery tools, it has played no role so far in constitutive discovery. The main reason is that constitution is commonly assumed to be characterized by (non-reductive) supervenience (see, e.g., Glennan [1996], pp. 61–2, and Eronen [2011], ch. 11), which generates deterministic dependencies: a complete set of constituents forms a supervenience base and thus a (modally) sufficient condition of a phenomenon, to the effect that there cannot be a change in the phenomenon without a change in its constituents. By contrast, as indicated in Section 2, PC is normally considered to be applicable to (pseudo)indeterministic data only.

To further clarify the difference between indeterministic and deterministic dependencies, consider the mechanism operating in an amplifier. Let $\mathcal{G}$ represent the phenomenon of gain, or absolute total voltage increase, of an amplifier subject to a voltage input $\mathcal{I}$. Amplifiers are built by assembling active elements, usually transistors, in a circuit. We assume that the amplifier in question is a two-stage amplifier, such that the signal received by a first transistor is amplified and fed to a second transistor, which further amplifies it. Let $\mathcal{A}$ and $\mathcal{B}$ be the transistors' absolute individual gains. Then, the amplifier's overall gain in response to any given input $\mathcal{I} = i$ is some indeterministic function $\mathcal{G} = r_{\mathcal{G}} i - i + \epsilon_{\mathcal{G}}$, where $r_{\mathcal{G}}$ indicates the amplifiers' amplification ratio and $\epsilon_{\mathcal{G}}$ is a noise term. For instance, if $\mathcal{I} = 2$ volts and the amplification ratio is 8, then the overall gain is $\mathcal{G} = 8 \times 2 - 2 + \epsilon_{\mathcal{G}}$ volts, where 14

(i.e., $8 \times 2 - 2$) volts and $\epsilon_{\mathcal{G}}$ volts, respectively, are $\mathcal{G}$'s deterministic and non-deterministic components. Analogously, the transistors' gains are also given by indeterministic functions, namely $\mathcal{A} = r_{\mathcal{A}} i - i + \epsilon_{\mathcal{A}}$ volts and $\mathcal{B} = r_{\mathcal{B}} i - i + \epsilon_{\mathcal{B}}$ volts. Assume that the first transistor amplifies by a ratio of 2, and the second amplifies by a ratio of 4.[3] Then, when subject to an input $\mathcal{I} = 2$ volts, the first transistor amplifies that signal by a gain of $2 \times 2 - 2 + \epsilon_{\mathcal{A}}$ volts; and the second transistor receives the amplified signal $(2+2+\epsilon_{\mathcal{A}})$ and amplifies it further by a gain of $4 \times (2+2+\epsilon_{\mathcal{A}}) - (2+2+\epsilon_{\mathcal{A}}) + \epsilon_{\mathcal{B}}$ volts. By contrast, the relation between overall gain $\mathcal{G}$ on the one hand, and the transistors' individual gains $\mathcal{A}$ and $\mathcal{B}$ on the other hand, is not indeterministic but deterministic: $\mathcal{G}$ is simply the sum of $\mathcal{A}$ and $\mathcal{B}$, meaning that $\mathcal{A}$ and $\mathcal{B}$ determine $\mathcal{G}$, such that whatever noisy component is present in $\mathcal{G}$, it is inherited from, and fully accounted for by, the noise in $\mathcal{A}$ and $\mathcal{B}$. More precisely, supervenience entails that $r_{\mathcal{G}} i - i + \epsilon_{\mathcal{G}} = r_{\mathcal{B}}(r_{\mathcal{A}} i + \epsilon_{\mathcal{A}}) - i + \epsilon_{\mathcal{B}}$. When $\mathcal{I} = 2$ volts, $8 \times 2 - 2 + \epsilon_{\mathcal{G}} = 4 \times (2+2+\epsilon_{\mathcal{A}}) - 2 + \epsilon_{\mathcal{B}}$, that is, $\epsilon_{\mathcal{G}} = 4\epsilon_{\mathcal{A}} + \epsilon_{\mathcal{B}}$.

Notwithstanding the frequency of CFC violations under determinism, (Gebharter [2017b], pp. 2652–54) has—surprisingly—argued that constitution satisfies the same axioms that PC assumes for causation. More specifically, he contends that the screening-off behaviour of complete sets of constituents (i.e., sets comprising a phenomenon's complete supervenience base) is analogous to that of deterministic direct causes and that the screening-off behaviour of incomplete sets is analogous to that of indeterministic direct causes. From that, he infers that constitutive relations can be represented by causal BNs and that, with some restrictions, PC is directly applicable to variable sets featuring both constitutive and causal relations, such that the uncovered dependencies can then be grouped into causal and constitutive dependencies by using knowledge of spatiotemporal overlap (i.e., parthood relations) between instances of variables. In short, Gebharter claims that the PC algorithm can perform causal and constitutive discovery in one go.

Given the well-known problems determinism creates for BN axioms, the natural conclusion to draw from Gebharter's finding that constitution behaves like deterministic direct causation would be that BNs are incapable of representing systems featuring constitutive relations, just as they incapable of representing systems featuring deterministic causal relations, and—*a fortiori*—that PC is inapplicable to systems featuring constitutive relations. Aware that his proposal raises severe questions, Gebharter discusses two approaches to reconcile the deterministic nature of constitution with BN axioms (cf. Gebharter [2017b], pp. 2661–62):

(A) Only apply PC to incomplete constitutive sets, which do not form complete supervenience bases and, hence, do not generate deterministic dependencies in the first place;

(B) Allow for deterministic dependencies but only apply PC to systems featuring no more

[3] This yields the amplifier's overall amplification ratio of 8 because a serial amplifier's amplification ratio is the product of its transistors' amplification ratios.

than two mechanistic levels.

Approach (A) amounts to testing for determinism prior to a BN analysis (by, e.g., performing a multicollinearity test) and, if that test is positive, abstaining from applying PC. A variable set **V** featuring constitutive relations will only be free of deterministic dependencies provided that no phenomenon in **V** has a complete set of constituents in **V**. As constitution, according to Gebharter, technically behaves like causation, missing constituents are on a par with missing causes of the phenomenon. Since constituents typically are not only relevant for the phenomenon but, on separate paths, also for other (micro-level) variables in **V**, it follows that missing constituents amount to missing common causes of two (or more) variables in **V**. If such common causes are missing, Sufficiency requires that they be fixed to the same value for all units in the population. However, as they are constituents of a phenomenon in **V** and, thereby, in the latter's supervenience base, they cannot be fixed, on pain of restraining the free variation of the phenomenon and, in consequence, of inducing spurious correlations between the phenomenon and other variables, in violation of CMC. In sum, missing constituents that are not fixed induce a violation of causal Sufficiency, in which case CMC tends to be violated, too (Gebharter [2017b], p. 2660). Thus, adopting approach (A) in an attempt to avoid CFC violations generates frequent CMC violations, which is why Gebharter concludes that (A) fails to reconcile the deterministic nature of constitution with the BN axioms. To justifiably assume CMC, **V** should contain complete constitutive sets, meaning that data over **V** should feature deterministic dependencies.

This leaves us with approach (B), which Gebharter indeed advances as a solution to the problems prompted by the deterministic nature of constitution (Gebharter [2017b], p. 2662). In the previous section, we have seen that chains of at least three deterministically related variables are a paradigmatic type of structure generating CFC violations. Without argument, Gebharter takes such chains to be the source of all CFC violations induced by determinism. Accordingly, he stipulates that PC only be applied to mechanistic systems with no more than two levels, which excludes deterministic chains.

To understand what exactly this two-level restriction entails, we need to clarify what Gebharter means by a level or by the predicate '... is at a different level from ...'. Unfortunately, he does not provide an explicit definition of this notion. In the mechanistic literature, the ordering among levels is often spelled out by means of the notion of constitution: $\mathcal{X}$ is at a lower level than $\mathcal{Y}$ if and only if $\mathcal{X}$ constitutes $\mathcal{Y}$ (see, e.g., Craver [2007], p. 189). However, in the context of constitutive discovery, where information about constitution is wanting, one cannot draw on constitution to fix the level ordering. That is, as the two-level restriction is intended to facilitate the inference to constitution, constitution cannot be presupposed in order to spell out the two-level restriction, meaning that the notion of a level cannot be understood along the lines of Craver ([2007], p. 189). But what can be presupposed in contexts of constitutive discovery is parthood information, which, we submit, is the next best conceptual basis to clarify Gebharter's two-level restriction.

Mechanists typically assume clarity on spatiotemporal parthood relations. That is, for any two variables $\mathcal{X}$ and $\mathcal{Y}$, they assume clarity on whether $\mathcal{X}$ and $\mathcal{Y}$ stand in a relation of proper parthood, that is, on whether the instances of $\mathcal{X}$ occupy a spacetime region strictly contained within the instances of $\mathcal{Y}$, or vice versa. For simplicity, we subsequently talk about variables standing in a parthood relation, even though, strictly speaking, only their instances may stand in such a relation. In the vein of Eronen ([2013], p. 1047), we define a variable $\mathcal{X}$ to be a direct proper part of a variable $\mathcal{Y}$ if and only if $\mathcal{X}$ is a proper part of $\mathcal{Y}$ and there does not exist another variable $\mathcal{Z}$ that is a proper part of $\mathcal{Y}$ such that $\mathcal{X}$ is a proper part of $\mathcal{Z}$. This yields an order of direct proper parthood, which can be used to locally distinguish spatiotemporal levels of a mechanistic phenomenon $\mathcal{Y}$: $\mathcal{Y}$ is on the top level; the variables representing direct proper parts of $\mathcal{Y}$ are on the next lower level; then come the variables representing direct proper parts of the direct proper parts of $\mathcal{Y}$, and so on.

Against that background, Gebharter's two-level restriction can now be said to stipulate that a variable set **V** constitutively analysed by PC contains variables from no more than two spatiotemporal levels. That is, **V** may include a phenomenon and its parts on one lower level in the spatiotemporal level hierarchy, but no parts of parts of it. Or differently, **V** must not contain any triple of variables $\langle \mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3 \rangle$ such that $\mathcal{V}_1$ is a part of $\mathcal{V}_2$, which is a part of $\mathcal{V}_3$. This restriction does not presuppose clarity on constitutive relations, since—to reiterate—not all parts are constituents. The restriction excludes that **V** contains constituents of constituents of a phenomenon, which, in turn, excludes that **V** contains deterministic chains inducing violations of CFC. Gebharter believes that it is thereby ensured that deterministic dependencies do not conflict with CFC more frequently than indeterministic dependencies and, hence, that CFC is justifiably assumable even for mixed variable sets featuring both phenomena and constituents—rendering PC applicable for the purpose of constitutive discovery.

The following sections critically review Gebharter's proposal.

4 Markov Violations Due to the Two-level Restriction

Before the next section discusses whether Gebharter's two-level restriction really renders CFC justifiably assumable for a variable set **V** comprising a complete set of constituents for every phenomenon in **V**, we want to point out that satisfying the two-level restriction for **V** amounts to excluding certain variables from **V**. This is bound to conflict with Sufficiency's call for includingall direct common causes or fixing them to constant values. If there exists a direct common cause of two variables in **V** that is on a third level with respect to two other variables in **V**, the two-level restriction requires it to be excluded from **V**. But if that direct common cause cannot be fixed to a constant value when the other variables in **V** vary, excluding it from **V** violates Sufficiency and, *a fortiori*, CMC.

To make this problem concrete, consider the structure in Figure 1, where $\mathcal{V}_1$ denotes a

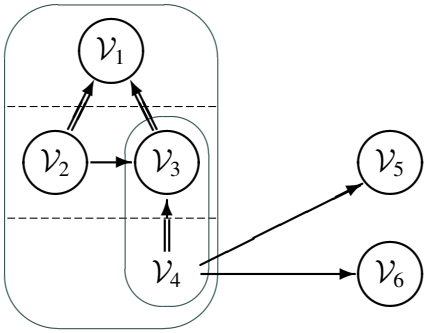

Figure 1: A hypothetical structure where arrows denote causation and double arrows constitution. Ovals contain variables in parthood relations, and dashed lines separate upper-level from lower-level variables. Latent variables ($\mathcal{V}_4$) are not inside circles.

phenomenon with two causally connected constituents $\mathcal{V}_2$ and $\mathcal{V}_3$. $\mathcal{V}_3$ itself has a constituent, $\mathcal{V}_4$, which is a common cause of two downstream effects $\mathcal{V}_5$ and $\mathcal{V}_6$. Subject to the two-level restriction, $\mathcal{V}_4$ must be excluded from the analysis when investigating the interplay between the variables in $\mathbf{V} = \{\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3, \mathcal{V}_5, \mathcal{V}_6\}$. But if $\mathcal{V}_4$ remains latent, Sufficiency requires that it be fixed to a constant value in all units of the population. This, however, cannot be done in a study scrutinizing the structure in Figure 1, which requires free variability of all involved variables. Since $\mathcal{V}_4$ is in the supervenience base of $\mathcal{V}_3$, fixing it would suppress the free variation of $\mathcal{V}_3$, giving rise to spurious correlations involving $\mathcal{V}_3$. In other words, excluding $\mathcal{V}_4$ from the analysed variable set violates Sufficiency, to the effect that the spurious correlations between $\mathcal{V}_5$ and $\mathcal{V}_6$ as well between $\mathcal{V}_3$ (and possibly $\mathcal{V}_1$) and $\mathcal{V}_5$/$\mathcal{V}_6$ cannot be screened off, in violation of CMC.

One might contend that CMC could be restored by, instead of $\mathcal{V}_4$, including suitable ancestors of $\mathcal{V}_4$ outside of the spacetime region occupied by $\mathcal{V}_1$, which *ipso facto* are also common causes of $\mathcal{V}_3$, $\mathcal{V}_5$, and $\mathcal{V}_6$. However, ancestors of $\mathcal{V}_4$ are not direct common causes of $\mathcal{V}_3$ and $\mathcal{V}_5$. They contain the same information about $\mathcal{V}_4$'s descendants as $\mathcal{V}_4$ itself only if they raise the probability of $\mathcal{V}_4$ as close to 1 as possible. If that is the case, however, further deterministic dependencies would result in addition to the ones induced by constitution, which would further raise the chances of CFC violations.

Of course, whether the two-level restriction gives rise to CMC violations crucially hinges on the particularities of an analysed structure. The problem only obtains in structures such that a part of a part is a non-fixable common cause of two variables in $\mathbf{V}$. We have no reason to assume that structures of this type are rare, but, at the same time, do not want to contend that they are particularly frequent. What we want to insist on, however, is that Gebharter is wrong in believing that once a complete set of constituents of every phenomenon in $\mathbf{V}$ is contained in $\mathbf{V}$ the remaining problem of how to ensure Sufficiency 'is just the general problem of how to guarantee for causal sufficiency, which all causal modeling approaches have to face' (Gebharter [2017b], p. 2661). Gebharter's two-level restriction, which he introduces to maintain CFC despite the presence of deterministic dependencies, is not without consequences on the difficulty of ensuring Sufficiency and CMC. Establishing the satisfac-

tion of Sufficiency and CMC is considerably more difficult in the case of mixed variable sets comprising variables from no more than two levels than it is in the case of ordinary variable sets analysed in causal modelling.

5 Extensive Faithfulness Violations

Now we turn to the question whether mechanistic systems on no more than two levels can justifiably be assumed to comply with CFC. Section 3 has shown that, to satisfy CMC, mechanistic systems must be analysed relative to variable sets $\mathbf{V}$ with complete sets of constituents $\mathbf{C}$. Subject to the supervenience of phenomena on their constituents, every such set $\mathbf{C}$ determines its corresponding phenomenon. This universal bottom-up determination yields that every phenomenon is screened off from all other variables—whether in $\mathbf{V}$ or not. The reason is that determination is monotonic: for any arbitrary variable $\mathcal{V}_i$, if $\mathbf{C}$ determines $\mathcal{V}_1$, then $\mathbf{C} \wedge \mathcal{V}_i$ also determines $\mathcal{V}_1$. If $\Pr(\mathcal{V}_1|\mathbf{C}) = 1$, then $\Pr(\mathcal{V}_1|\mathbf{C} \wedge \mathcal{V}_i) = 1$, meaning that $\mathbf{C}$ screens off $\mathcal{V}_1$ from any variable $\mathcal{V}_i$. CFC is only satisfied in such contexts if the conditional independencies between the phenomena and all their non-constituents are entailed by the true graphs, meaning that all macro phenomena in fact are both uncaused (they only have incoming arrows from their supervenience-base variables) and causally inert, that is, causally isolated.

To illustrate, reconsider our amplifier example from Section 3 and let the analysed variable set be $\mathbf{G} = \{\mathcal{I}, \mathcal{G}, \mathcal{S}, \mathcal{A}, \mathcal{B}\}$, where $\mathcal{I}$ (the amplifier's input), $\mathcal{G}$ (the amplifier's overall gain), $\mathcal{A}$ (the first transistor's absolute gain), and $\mathcal{B}$ (the second transistor's absolute gain) are complemented by $\mathcal{S}$, which denotes, say, the distortion of the signal received by a loudspeaker. Since $\mathcal{A}$ and $\mathcal{B}$ determine $\mathcal{G}$ (from the bottom up), $\mathcal{A}$ and $\mathcal{B}$ screen off $\mathcal{G}$ from $\mathcal{I}$ and $\mathcal{S}$, or formally $\mathcal{I}, \mathcal{S} \perp\!\!\!\perp \mathcal{G} \,|\, \mathcal{A}, \mathcal{B}$.[4] These conditional independencies only comply with CFC if the amplifier's overall gain *de facto* is neither caused by its input nor a cause of the distortion in the loudspeaker. More generally, to satisfy CFC, the amplifier's overall gain must be assumed to be causally isolated from the rest of the universe.

That, of course, is a pill impossible to swallow for most mechanists, that is, the addressees of Gebharter's proposal, as they tend to be non-reductive physicalists who endorse the existence of macro-level causation.[5] They will thus reject the causal isolation of all phenomena and, consequently, interpret the conditional independencies between phenomena and all non-constituents in contexts featuring complete sets of constituents as CFC violations that obtain even in two-level systems.

To avoid CFC violations, Gebharter ([2017a]), in turn, rejects non-reductive physicalism

[4] By contrast, $\mathcal{A}$ does not screen off $\mathcal{I}$ and $\mathcal{B}$. When holding the absolute gain of the first transistor fixed, $\mathcal{I}$ still makes a difference to the absolute gain of the second transistor. For the same reason, $\mathcal{B}$ does not screen off $\mathcal{A}$ and $\mathcal{S}$, and $\mathcal{A}$ and $\mathcal{B}$ do not screen off $\mathcal{I}$ and $\mathcal{S}$.

[5] In fact, we are not aware of a single proponent of the mechanistic framework who would endorse the causal isolation of macro phenomena.

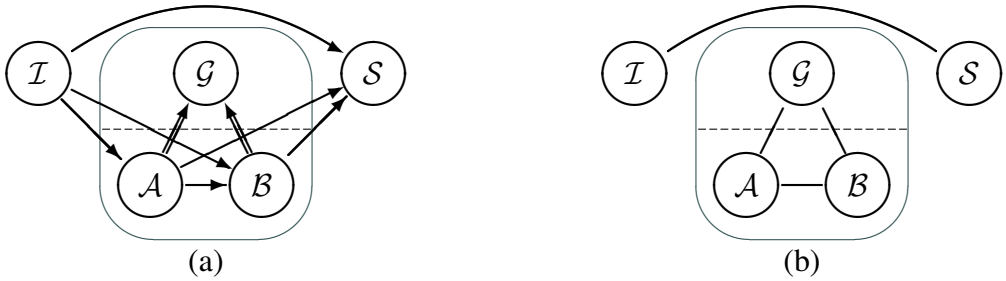

Figure 2: Graph (a) is the true structure of a two-stage amplifier mechanism over $\mathbf{G} = \{\mathcal{I}, \mathcal{G}, \mathcal{S}, \mathcal{A}, \mathcal{B}\}$ for an epiphenomenalist*. Graph (b) results from applying PC to the true conditional (in)dependencies over $\mathbf{G}$, where $\mathcal{G}$ is a deterministic function of $\mathcal{A}$ and $\mathcal{B}$.

and endorses a radical form of macro-level epiphenomenalism, call it epiphenomenalism*, namely the view that non-fundamental properties are not only causally inert (as entailed by standard epiphenomenalism) but also uncaused.[6] More concretely, according to epiphenomenalism*, the true graph for our amplifier example is the one in Figure 2a. Against that backdrop, the fact that $\mathcal{G}$ is screened off from $\mathcal{I}$ and $\mathcal{S}$ by $\mathcal{A}$ and $\mathcal{B}$ follows from CMC applied to the true graph and, hence, does not violate CFC. Clearly though, this manœuvre not only clashes with the standard metaphysical commitments in the mechanistic literature but also with the scientific practice of those disciplines that are most interested in constitution, such as the social and biomedical sciences, which routinely engage in investigating causal relations among macro variables and, hence, do not commit to epiphenomenalism*.

Worse yet, in addition to bottom-up determination, mechanistic systems with no more than two levels may also feature top-down determination, to the effect that not only phenomena are screened off from all incoming and outgoing influences, but also constituents can be screened off in this way. This problem is best introduced by reconsidering the amplifier example. The amplifier's absolute overall gain $\mathcal{G}$ is the sum of its constituents $\mathcal{A}$ and $\mathcal{B}$. The function of addition, however, is reversible: it not only holds that $\mathcal{G}$ is determined by $\mathcal{A}$ and $\mathcal{B}$, but also that $\mathcal{A}$ is determined by $\mathcal{G}$ and $\mathcal{B}$ (e.g., $\mathcal{G} = 14 \wedge \mathcal{B} = 12$ determines $\mathcal{A} = 2$) and that $\mathcal{B}$ is determined by $\mathcal{G}$ and $\mathcal{A}$ (e.g., $\mathcal{G} = 14 \wedge \mathcal{A} = 2$ determines $\mathcal{B} = 12$). Hence, every variable in $\mathbf{M} = \{\mathcal{G}, \mathcal{A}, \mathcal{B}\}$ is screened off from $\mathcal{I}$ and $\mathcal{S}$ by the other two elements of $\mathbf{M}$.

If PC is applied to oracle information on conditional (in)dependencies in $\mathbf{G}$, all edges connecting the variables in $\mathbf{M}$ to the variables in $\mathbf{G} \setminus \mathbf{M}$ will be removed, resulting in the graph skeleton in Figure 2b. This graph is non-Markovian because the pairs $\langle \mathcal{I}, \mathcal{A} \rangle$, $\langle \mathcal{I}, \mathcal{B} \rangle$, $\langle \mathcal{A}, \mathcal{S} \rangle$, $\langle \mathcal{B}, \mathcal{S} \rangle$ are unconnected even though these variables are pairwise unconditionally dependent. Under the assumption that CMC is satisfied, Figure 2b cannot amount to the skeleton of the true graph because too many edges have been eliminated. Moreover, since no Markovian

[6] Gebharter insists ([2017b], p. 2660) that macro variables may still be involved in 'inefficient' (or 'unproductive') causal relations, that is, causal relations that do not manifest themselves in difference-making patterns in data and, hence, are undetectable by methods of causal data analysis. Hence, Gebharter endorses epiphenomenalism* with respect to efficient causation only. But admitting inefficient relations in a formalism that takes difference-making to be necessary for causation is unfounded and ad hoc.

graph over **G** exists that entails all the independencies depicted in Figure 2b, this constitutes a so-called detectable violation of CFC (Zhang and Spirtes [2016], p. 252), namely a CFC violation ensuing from the fact that the data cannot possibly be modelled in compliance with BN axioms.[7] No metaphysical background assumption—whether epiphenomenalism* or else—could ever reconcile the independencies in Figure 2b with CFC. The only remaining conclusion is that CFC is violated in this two-level mechanism and, a fortiori, that PC is inapplicable to our amplifier system.

The possibility of top-down determination shows that not even the idiosyncratic metaphysical background of epiphenomenalism* suffices to secure the applicability of PC to mixed variable sets complying with Gebharter's two-level restriction. The crucial follow-up question now becomes how widespread top-down determination is. It is clearly not limited to amplifier gains or even to phenomena whose values are the sum of their constituents. It obtains whenever the relation between phenomena and constituents is regulated by an aggregation function with the following reversibility property: a function $y = f(x_1, \ldots, x_n)$ is reversible if and only if all of its inputs x_i are determined by its output y in conjunction with all of its other inputs apart from x_i, or formally, if and only if for all i, $1 \leq i \leq n$, $x_i = f^{-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, y)$. Examples of functions for which reversibility holds are linear functions, the product of non-zero values, exponentiation of positive integers, the sum of squares, many Boolean functions, or functions used in information coding, storage, and encryption (which are explicitly exploited for their reversibility).

To provide another example, consider the phenomenon of voting by a show of hands. Casting a vote, $\mathcal{W} = 1$, can be constituted by a raise of either the left hand, $\mathcal{L} = 1$, or of the right hand, $\mathcal{R} = 1$ (but raising both hands is invalid); or formally, $\mathcal{W} = 1 \leftrightarrow (\mathcal{L} = 1 \wedge$ R=0) $\vee$ $(\mathcal{L} = 0 \wedge \mathcal{R} = 1)$. This system of binary variables does not only feature bottom-up determination but also top-down determination: any of the four possible value configurations of $\{\mathcal{W}, \mathcal{L}\}$ and of $\{\mathcal{W}, \mathcal{R}\}$ determine the value of $\mathcal{R}$ and $\mathcal{L}$, respectively.[8] Hence, not only the phenomenon of voting but also the hand raisings are screened off from all variables outside of that system. But hand raisings, for example, have causes in the motor cortex and effects in air displacement, meaning that outside variables can *de facto* causally interact with $\mathcal{R}$ and $\mathcal{L}$. That these outside variables can be screened off from $\mathcal{R}$ and $\mathcal{L}$ in mixed variable sets comprising complete sets of constituents, therefore, violates CFC.

These considerations suffice to establish that, contrary to what Gebharter envisages in approach (B), CFC violations in (deterministic) mechanistic systems comprising only two levels are not rare but widespread—unlike CFC violations in (pseudoindeterministic) causal

[7] We thank an anonymous referee for pointing this out to us. The detectability of the CFC violation would render the (conservative) PC algorithm applicable if it was only the so-called Orientation Faithfulness component of CFC that was violated (Zhang and Spirtes [2016], pp. 254–55). What is violated here, however, is the Adjacency Faithfulness component.

[8] To illustrate for $\{\mathcal{W}, \mathcal{L}\}$ and $\mathcal{R}$: $\mathcal{W}=0 \wedge \mathcal{R}=0 \rightarrow \mathcal{L}=0$; $\mathcal{W}=0 \wedge \mathcal{R}=1 \rightarrow \mathcal{L}=1$; $\mathcal{W}=1 \wedge \mathcal{R}=0 \rightarrow \mathcal{L}=1$; and $\mathcal{W}=1 \wedge \mathcal{R}=1 \rightarrow \mathcal{L}=0$.

systems. The two-level restriction does not warrant the justifiable assumability of CFC.

A possible response might be to further restrict the applicability of PC to mechanisms regulated by non-reversible aggregation functions. Paradigmatic non-reversible functions are periodic functions, products of zero, or the maximum and minimum functions. If a phenomenon is aggregated from its constituents by a non-reversible function, it does not hold for every constituent that its values are determined by the phenomenon in conjunction with all other constituents, that is, top-down determination does not obtain. However, such an approach would differ in a crucial respect from Gebharter's original restriction to two-level systems in (B). A variable set **V** can be ensured to comply with the two-level restriction by imposing that **V** does not contain a triple $\langle \mathcal{V}_i, \mathcal{V}_j, \mathcal{V}_k \rangle$ such that $\mathcal{V}_i$ is a spatiotemporal part of $\mathcal{V}_j$ and $\mathcal{V}_j$ is a part of $\mathcal{V}_k$. While identifying spatiotemporal parthood relations—clarity on which is generally assumed in the mechanistic literature—is undoubtedly difficult, it does not presuppose clarity on constitutive relations. In consequence, that **V** satisfies the two-level restriction can be established independently of clarity on the constitutive relations obtaining among the elements of **V**. The same does not hold for a restriction to admissible aggregation functions. It is unclear how it could be established independently of clarity on the identity of the constituents that a phenomenon is aggregated from its constituents in **V** by a certain type of (non-reversible) function. What type of function regulates the interplay between phenomena and constituents can only be determined after the constituents have been identified. The latter, however, is exactly the purpose of Gebharter's procedure. Hence, an attempt to avoid CFC violations resulting from top-down determination by restricting the procedure's applicability to systems with non-reversible aggregation functions would render that procedure circular.

Nonetheless, let us assume for the sake of the argument that there are types of mechanistic systems for which the nature of the aggregation function is known even in the absence of clarity on the constituents. The applicability of Gebharter's proposal could thus be confined to mechanisms known to have a non-reversible aggregation function. To show that not even such a restriction would ensure compliance with CFC, we modify the voting example such that a vote also counts as validly cast ($\mathcal{W}=1$) if both hands are raised ($\mathcal{L}=1 \wedge \mathcal{R}=1$). The relation between the phenomenon $\mathcal{W}$ and its constituents $\mathcal{L}$ and $\mathcal{R}$ shall hence be regulated by the non-reversible function of inclusive disjunction (i.e., maximum): $\mathcal{W}=1 \leftrightarrow \mathcal{L}=1 \vee \mathcal{R}=1$ (i.e., $\mathcal{W} = max(\mathcal{L}, \mathcal{R})$). While we still get bottom-up determination from this system, that is, every value configuration of $\{\mathcal{L}, \mathcal{R}\}$ determines a value of $\mathcal{W}$, we no longer get top-down determination. Not every value configuration of $\{\mathcal{W}, \mathcal{R}\}$ and $\{\mathcal{W}, \mathcal{L}\}$ determines a value of $\mathcal{L}$ and $\mathcal{R}$, respectively. For example, if $\mathcal{W}=1$ and $\mathcal{L}=1$, it is not determined whether $\mathcal{R}$ takes the value 0 or 1, as both values are possible.

To decide whether Gebharter's procedure is reliably applicable to structures for which top-down determination can be non-circularly excluded, we embed this non-reversible voting mechanism in a simple causal context. Let $\mathcal{M}$ be a variable representing the cause of the

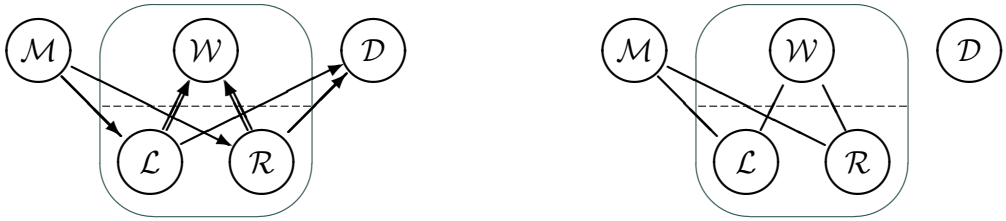

Figure 3: Voting with a non-reversible aggregation function. (a) is the true graph over $\mathbf{O} = \{\mathcal{M}, \mathcal{L}, \mathcal{R}, \mathcal{W}, \mathcal{D}\}$ under epiphenomenalism* (where double arrows are constitutive). Graph (b) results from applying PC to the true conditional (in)dependencies over $\mathbf{O}$.

hand raising in the voter's motor cortex, and let $\mathcal{D}$ represent the ultimate decision taken by the vote. Let us moreover grant Gebharter that epiphenomenalism* holds. It follows that the true structure over $\mathbf{O} = \{\mathcal{M}, \mathcal{L}, \mathcal{R}, \mathcal{W}, \mathcal{D}\}$ is the one in graph (a) of Figure 3. Contrary to constitutive arrows, causal arrows shall again be indeterministic. In that system, $\mathcal{L}$ and $\mathcal{R}$ cannot be screened off from their cause $\mathcal{M}$ by the other variables in $\mathbf{O}$. However, since $\mathcal{W}$ is a deterministic function of $\mathcal{L}$ and $\mathcal{R}$, and $\mathcal{D}$ can be expressed as a probabilistic function of $\mathcal{W}$, $\mathcal{W}$ encodes all the information on $\mathcal{L}$ and $\mathcal{R}$ relevant for the probability of $\mathcal{D}$. All that matters for the decision is whether at least one hand was raised; whether it was the left or the right is irrelevant. Hence, given the value of $\mathcal{W}$ additional information about $\mathcal{L}$ or $\mathcal{R}$ has no bearing on the probability of $\mathcal{D}$. Or formally, $\mathcal{D} \perp\!\!\!\perp \mathcal{L}, \mathcal{R} \,|\, \mathcal{W}$. Even without top-down determination, $\mathcal{W}$ screens off the hand raisings from the resulting decision. If PC is applied to oracle information on conditional (in)dependencies in $\mathbf{O}$, it will detach $\mathcal{D}$ from the voting mechanism, as shown in Figure 3b. Just as Figure 2b, Figure 3b is non-Markovian because the pairs $\langle \mathcal{W}, \mathcal{D} \rangle$, $\langle \mathcal{L}, \mathcal{D} \rangle$, $\langle \mathcal{R}, \mathcal{D} \rangle$ are unconnected despite the fact that they are unconditionally dependent. Since no Markovian graph exists, which is faithful to the (in)dependencies among the variables in $\mathbf{O}$, CFC is again detectably violated, which, in turn, establishes PC's inapplicability—the two-level restriction (and epiphenomenalism*) notwithstanding. In sum, strengthening approach (B) by adding a restriction to certain types of aggregation functions is not a feasible option.

The question explored in this section must be answered in the negative: Mechanistic systems on no more than two levels cannot justifiably be assumed to comply with CFC. This confirms the received wisdom in the BN literature that variable sets comprising phenomena and their constituents are simply beyond the scope of warranted applicability of PC, which is limited to indeterministic data (cf. condition 3 in Spirtes, Glymour, and Scheines [2000], p. 351).

6 PCD Won't Save the Day

Given the problems deterministic data generate for PC, Glymour ([2007]) has proposed a variant of PC, called PCD, that is custom-built for variable sets featuring deterministic de-

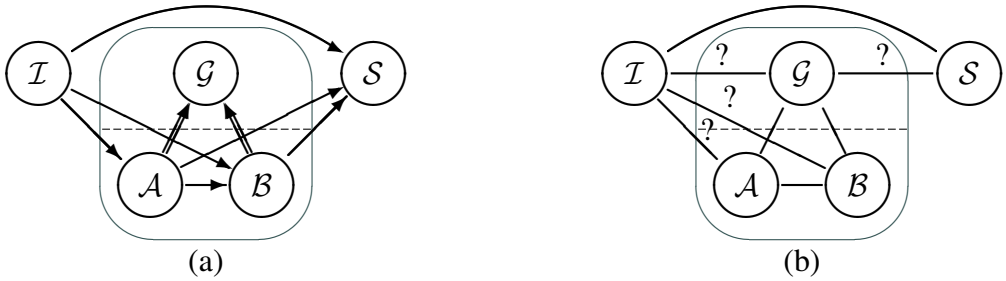

Figure 4: (a) is the true graph over **G**, and (b) the skeleton output by PCD applied to the true conditional (in)dependencies in **G**.

pendencies. Accordingly, this section investigates whether the principle behind Gebharter's proposal could be saved by implementing it with PCD instead of PC. PCD aims to make causal discovery insensitive to CFC violations induced by determinism. To this end, it operates like PC with one important exception. Unlike PC, PCD does not take screen-off relations involving maximal conditional probabilities of 1 to indicate the absence of causation. PCD only infers that two variables $\mathcal{V}_i$ and $\mathcal{V}_j$ are causally unrelated if they can be screened off with non-maximal conditional probabilities. If they can only be screened off with maximal probability, the output of PCD features an edge between $\mathcal{V}_i$ and $\mathcal{V}_j$ that is marked as 'uncertain' with a question mark (Glymour [2007], p. 236).

The first thing to note about the idea of replacing PC by PCD in Gebharter's procedure is that discovery by PCD is much less informative than by PC. While PC exploits conditional independencies of 1 to infer to (causal) irrelevance, PCD simply abstains from drawing any inference from such independencies. Furthermore, it is doubtful whether the assumptions required by PCD are any more justifiable when analysing mechanistic systems than the assumptions of PC—even though PCD's assumptions are clearly weaker than PC's. While applying PC requires assuming that all conditional independencies in the data faithfully reflect the true graph, applying PCD only requires assuming that the conditional independencies with probabilities lower than 1 are faithful to the true graph. But the version of the voting example with a non-reversible aggregation function (*max*) has shown that bottom-up determination may generate non-deterministic screen-off relations that do not follow from applying CMC to the true graph. The same happens in our amplifier example. Since the overall gain, $\mathcal{G}$, is the sum of the individual gains, $\mathcal{A}$ and $\mathcal{B}$, of the amplifier's two transistors, $\mathcal{G}$ encodes all the information on $\mathcal{A}$ and $\mathcal{B}$ that is relevant for the probability of distortion, $\mathcal{S}$. Accordingly, although $\mathcal{S}$ is not determined by any subset of $\mathbf{G} = \{\mathcal{I}, \mathcal{G}, \mathcal{S}, \mathcal{A}, \mathcal{B}\}$, it is screened off from $\mathcal{A}$ and $\mathcal{B}$ by the conjunction of the input to the amplifier, $\mathcal{I}$, and $\mathcal{G}$: if we know the values of $\mathcal{I}$ and $\mathcal{G}$, additional information on $\mathcal{A}$ and $\mathcal{B}$ has no bearing on the probability of $\mathcal{S}$, or formally, $\mathcal{S} \perp\!\!\!\perp \mathcal{A}, \mathcal{B} \,|\, \mathcal{I}, \mathcal{G}$. These conditional independencies obtain despite $\mathcal{I}$ and $\mathcal{G}$ not raising the probability of $\mathcal{S}$ to 1 and, hence, should be faithful to the true graph, if PCD is applied to data on our amplifier. However, they are not.

If PCD is applied to oracle information on conditional (in)dependencies among the variables in **G**, its output has the skeleton in Figure 4b. Here, CFC is violated because the edges

corresponding to the pairs $\langle \mathcal{A}, \mathcal{S} \rangle$ and $\langle \mathcal{B}, \mathcal{S} \rangle$ are missing, even though $\mathcal{A}$ and $\mathcal{B}$ are causes of $\mathcal{S}$ (cf. Figure 4a). Moreover, contrary to the graphs in Figures 2b and 3b, this graph is Markovian, as it preserves connections corresponding to all unconditional dependencies. In particular, the pairs $\langle \mathcal{A}, \mathcal{S} \rangle$ and $\langle \mathcal{B}, \mathcal{S} \rangle$ are connected—via $\mathcal{G}$. This means that, differently from the CFC violations incurred by PC, this CFC violation is non-detectable.

Clearly, these (non-deterministic) CFC violations do not hinge on the particularities of the voting or the amplifier example. If a set of variables $\mathbf{D}$ determines a variable $\mathcal{V}_i$, it easily happens that $\mathcal{V}_i$ encodes all the information on $\mathbf{D}$ relevant to the probability of some downstream variable $\mathcal{V}_j$. In all such cases, $\mathcal{V}_i$ renders $\mathcal{V}_j$ conditionally independent of $\mathbf{D}$, even if the corresponding conditional probabilities are below 1. Undoubtedly, this is a frequent pattern in systems featuring phenomena and complete sets of their constituents. According to all metaphysical views that do not deny the causal efficacy of constituents, these (non-deterministic) conditional independencies violate the faithfulness standards of PCD and, thus, render the use of PCD unwarranted—again, despite compliance with the two-level restriction. Moreover, since these CFC violations are undetectable, the inapplicability of PCD will tend to go unnoticed. Consequently, PCD may be unjustifiably applied resulting in fallacious inferences. By contrast, the detectability of the CFC violations incurred by PC ensures that PC's inapplicability does not go unnoticed, thereby preventing fallacious inferences. In sum, PCD is an even less suitable tool for constitutive inference than PC.

7 False Positives

Recently, various studies (e.g., Zhang and Spirtes [2008], Zhalama, Zhang, and Mayer [2017]) have investigated to what degree CFC violations affect the actual output of PC, among other algorithms. These studies suggest that proper parts of PC's outputs can, under certain circumstances, be reliably interpreted causally despite CFC violations. More concretely, it is CFC's purpose to ensure that the absence of edges in PC's outputs can be interpreted in terms of the absence of causation. This interpretation is blocked if CFC is violated. However, the interpretation of present edges in terms of the presence of causation remains unaffected by CFC violations. So perhaps there is a case to be made that, when applied to mechanistic systems, PC can still reliably infer the presence of causal/constitutive dependence relations without incurring an unacceptable risk of committing false positives, even if it cannot reliably infer the absence of such relations, due to a severe risk of false negatives. If this holds up to scrutiny, Gebharter's approach could be used as a means to uncover the presence of constitutive and causal dependence relations in mixed variable sets, even if it does not reliably exhibit their absence.

To investigate that question we set up two benchmark experiments; the first running PC with the frequently used parametric independence test Fisher's Z, and the second running it with a promising non-parametric independence test called RCIT (Strobl, Zhang, and

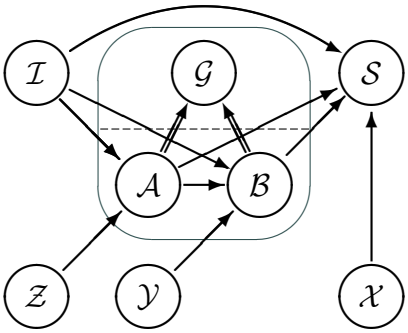

Figure 5: PC-friendly expansion of the structure in Figure 2a over $\mathbf{G}^* = \mathbf{G} \cup \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$.

Visweswaran [2018]). Each experiment consists of two series of inverse search trials testing the reliability of PC's analysis of data simulated from a Gaussian data-generating structure (in the sense of Edwards [2000], ch. 3) whose core has the form of the mechanism behind our amplifier example. In both experiments, the first trial series has the objective to determine the false positive ratios among unoriented and oriented edges issued by PC when applied to data featuring deterministic dependencies, and the second series to determine the ratio among these false positives ascribable to determinism. For that purpose, the only difference between the data-generating structures used for the two series is that the structure for the first series induces deterministic dependencies between certain variables whereas the one for the second series does not. More precisely, in both trial series, all variables are indeterministic, that is, sampled with (normally distributed) error terms, with the exception of variable $\mathcal{G}$, which is aggregated from variables $\mathcal{A}$ and $\mathcal{B}$ deterministically (i.e., without its own, independent error term) in the first series, and indeterministically (i.e., with its own error term) in the second.

The quality of PC's outputs is known to be sensitive to various factors, such as the existence of unshielded colliders, the sample size, the joint normality of the distribution or the linearity of the functional dependencies (see, e.g., Spirtes, Glymour, and Scheines [2000], p. 351). As deterministic dependencies induced by constitution shall be the only obstacle for PC, we ensure that the trials are otherwise favourable to PC. To this end, we do not directly simulate data from the amplifier structure in Figure 2a but expand it by adding three unshielded colliders, one on the transistor variables $\mathcal{A}$ and $\mathcal{B}$ each, and one on the distortion variable $\mathcal{S}$ (see Figure 5).

The false positive ratios in the two trial series will, of course, depend on what we take the true data-generating structure to be. In the second—purely causal—series, the true structure is straightforwardly obtained by a causal interpretation of the graph in Figure 5, according to which the edges $\mathcal{A} \longrightarrow \mathcal{G}$ and $\mathcal{B} \longrightarrow \mathcal{G}$ are causal and not constitutive. In case of the first—mechanistic—series, however, different background metaphysics disagree on what the true data-generating structure is, as we have seen in Section 5. In order to remain maximally charitable to Gebharter, we grant him his epiphenomenalism* and simulate the data in the first trial series such that the amplifier's input $\mathcal{I}$ appears in the equations generating the

values of $\mathcal{A}$ and $\mathcal{B}$, which, in turn, appear in the equations generating the values of $\mathcal{S}$. Yet, the variable representing the overall gain, $\mathcal{G}$, is neither simulated as a function of $\mathcal{I}$ nor is $\mathcal{S}$ simulated as a function of $\mathcal{G}$. In other words, we assume that the true graph over $\mathbf{G}^* = \{\mathcal{I}, \mathcal{G}, \mathcal{S}, \mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$ does not comprise causal arrows in and out of $\mathcal{G}$ but that all causal influence goes through the constituents $\mathcal{A}$ and $\mathcal{B}$. That is, the true structure shall be the one obtained from an epiphenomenalist* interpretation of the graph in Figure 5.

We conduct the trials in R using the PC implementation **pcalg** by Kalisch, Mächler, Colombo, Maathuis, and Bühlmann ([2012]). Replication scripts for the two experiments are available in the paper's Appendix. For all trial series, we simulate 2000 data sets from the structure in Figure 5 with a (large) sample size of 10 000 observations each. The first trial series (in both experiments), processes the first 1000 data sets, and the second the remaining 1000 data sets. We draw normally distributed values for all variables and for all (mutually independent) error terms, all being centred around 0 and having standard deviations of 1. All variables are related by linear functions. To avoid that our results are sensitive to any numeric elements of those linear functions, we randomly draw—for each of the 2000 data sets—non-zero numeric constants for exogenous variables and parameters for endogenous variables (both from the interval $[-5, 5] \setminus \{0\}$).

In the two trial series of the first experiment, we run the `pc()` function from the **pcalg** package using the independence test `gaussCItest` implementing Fisher's Z. Moreover, we apply the majority rule (`maj.rule = TRUE`) from (Colombo and Maathuis [2014]) that checks triples of variables for orientation ambiguities and we set the significance level to $\alpha = 0.05$. We are interested in false positive ratios for edges and orientations. In an individual trial, these ratios correspond to the number of unoriented/oriented edges contained in the output graph but not in the corresponding true graph of Figure 5, divided by the total number of edges in the output graph. We also report completeness (or recall) ratios, that is, the number of unoriented/oriented edges contained both in the output graph and the true graph divided by the total number of edges in the true graph. In addition, we exhibit the recovery rates for a number of oriented edges that are relevant for our ensuing discussion of the results. The bar chart in Figure 6 presents the results of the first experiment; it shows the means of all of the above ratios over the 1000 trials of the first series in black and of the 1000 trials of second series in gray.

We find a significant difference in false positive ratios. Under determinism, on average 25.0% of the edges and 29.7% of the orientations are false. Under indeterminism, those numbers go down to 8.2% and 11.6%, respectively. That is, $\mathcal{G}$ being a deterministic function of its constituents triples the false positive ratios. Under the conditions favourable to its performance, PC performs satisfactorily when it comes to identifying edges and orientations. But the presence of only one determined variable leads to 3 of 10 orientations being wrong, which is a performance hardly describable as satisfactory under otherwise ideal discovery conditions.

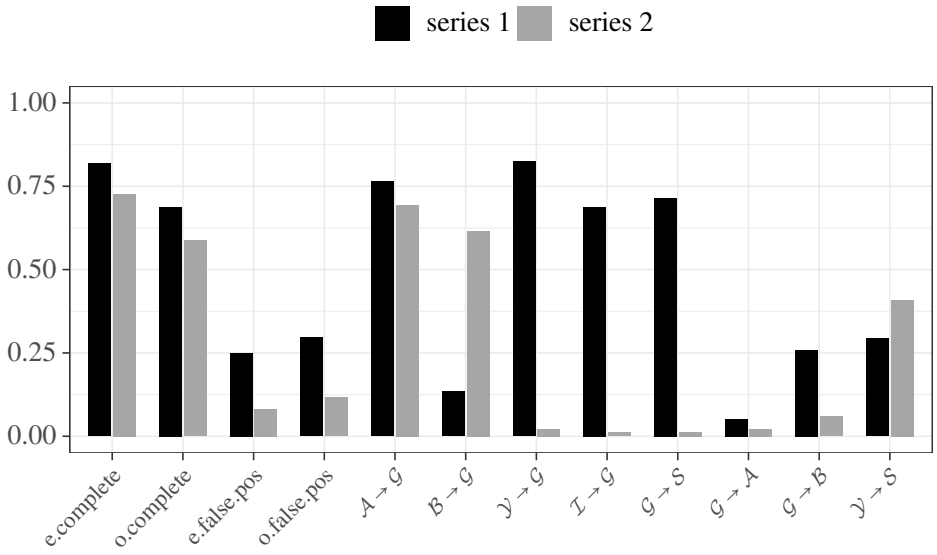

Figure 6: Results of the first benchmark experiment running PC with the parametric independence test Fisher's Z. Bar plots present completeness ratios for edges (e.complete) and orientations (o.complete); false positive ratios for edges (e.false.pos) and orientations (o.false.pos); and recovery rates for a number of interesting oriented edges.

One reason for the difference in false positive ratios is that PC finds less screen-off relations under determinism, resulting in more inferred adjacencies per trial overall. On average, PC issues 12 adjacencies per trial in the first series and only 8.7 in the second series. Some of these additional adjacencies are true, some of them not; some of them are correctly oriented, some of them not. But the difference in false positives between the two series is only to a small degree ascribable to the difference in the overall number of inferred adjacencies, as can be seen by comparing the decrease in completeness and false positive ratios between the two trial series. When $\mathcal{G}$ is a deterministic function (i.e., without error term) of $\mathcal{A}$ and $\mathcal{B}$, 81.8% of the edges and 68.8% of the orientations of the true graph are recovered on average. When $\mathcal{G}$ is sampled with an error term, those ratios go down to 72.7% and 58.7%, respectively. That is, the recovery rate of the true graph decreases by 11% for unoriented edges and by 20% for orientations. However, as we have seen above, the false positive rates decrease by over 60% when $\mathcal{G}$ is not determined by $\mathcal{A}$ and $\mathcal{B}$. That is, under indeterminism PC does not simply avoid mistakes because it infers less adjacencies overall, but because it hits the target more reliably. Or inversely put, determinism induces PC to mistake spurious for causal dependencies.

The culprit is Fisher's Z test, which is unreliable in detecting spurious independencies induced by determinism. The test measures partial correlations, and the concept of partial correlation (e.g., the correlation between X and Y given Z) is not well-defined when certain pairwise correlations are deterministic (e.g., when Z determines X or Y). With just one controlling variable Z, the partial correlation function, namely $\rho_{X,Y \cdot Z}$, is defined in terms of the pairwise correlations ρ_{XY}, ρ_{YZ}, and ρ_{XZ}, as follows:

$$\rho_{X,Y \cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1 - \rho_{XZ}^2} \cdot \sqrt{1 - \rho_{ZY}^2}}$$

When correlations are well defined, they take values over the closed interval $[-1, 1]$, where 0 indicates absence of correlation, -1 indicates full anticorrelation, and 1 indicates full correlation (Gentle [2013], p. 37). If Z determines X or Y, ρ_{XZ} or ρ_{ZY} are 1, such that one of the two squared roots, and thus the denominator, goes to 0, and the partial correlation function goes to infinity. It follows that conditional independencies due to determinism erroneously appear as maximal correlations.[9] This problem does not only affect Fisher's Z test, but any parametric test of conditional independence. In the family of graphical Gaussian models for continuous data, pairwise conditional independencies correspond to zero partial correlations (Edwards [2000], pp. 11, 36–7). In other words, under the assumption that all involved variables are multivariate Gaussian, conditional independence between any two variables holds if, and only if, their partial correlation is zero. All parametric independence tests try to reject a null hypothesis about (zero) partial correlation. Whenever PC is run with a parametric test, therefore, deterministic dependencies will induce false positives due to the unreliability of those tests in the presence of determinism.

In the first trial series, this problem is visible in the high output rates of false oriented edges in and out of $\mathcal{G}$. For example, $\mathcal{Y} \longrightarrow \mathcal{G}$ is issued in 82.4% of the trials, $\mathcal{I} \longrightarrow \mathcal{G}$ in 68.7%, and $\mathcal{G} \longrightarrow \mathcal{S}$ in 71.5%. In fact, none of these variables should be adjacent, because $\mathcal{A}$ and $\mathcal{B}$ determine $\mathcal{G}$ and, hence, screen $\mathcal{G}$ off from all other variables. But these independencies are not correctly recovered by Fisher's Z; and the resulting edges are even oriented by PC in the majority of the trials. These false positives disappear almost completely in the second series, where $\mathcal{Y} \longrightarrow \mathcal{G}$ is returned in negligible 2.2% of the trials, $\mathcal{I} \longrightarrow \mathcal{G}$ in 1.1%, and $\mathcal{G} \longrightarrow \mathcal{S}$ in 1.3%.

Before we investigate whether PC's performance under determinism can be improved by running it with a non-parametric independence test, we complete the discussion of the results from our first experiment. Figure 6 shows that, under determinism, PC correctly identifies the arrow $\mathcal{A} \longrightarrow \mathcal{G}$ in a remarkable 76.4% of the trials, which is significantly higher than the false positive rate. At the same time, the recovery rate for $\mathcal{B} \longrightarrow \mathcal{G}$ is only 13.5%, meaning

[9] Typical software implementations of Fisher's Z, such as `gaussCItest`, have in-built heuristics to avoid the explosion of the partial correlation function. As a result, independencies induced by determinism will often not appear as maximal but as ordinary non-maximal correlations. Moreover, there exist approaches for correcting extreme values of partial correlations by, for instance, shrinking the empirical correlations towards the identity matrix (Schäfer and Strimmer [2005]). When Fisher's Z is run on a thus shrunk correlation matrix (e.g., produced via `corpcor::pcor.shrink()`), it succeeds in reliably detecting independencies induced by determinism (we thank Marco Scutari for pointing this out; see the replication script for a concrete example). However, that works reliably only with a properly pre-determined λ value and it misleads PC into frequently inferring too many independencies, to the effect that ordinary causal dependencies are no longer properly recovered.

that the prospect of discovering that $\mathcal{B}$ is a constituent of $\mathcal{G}$ is less than half as high as the risk of inferring a false orientation. What is more, the edge connecting $\mathcal{B}$ and $\mathcal{G}$ is mis-oriented from $\mathcal{G}$ to $\mathcal{B}$ twice as much (25.9%) as it is oriented from $\mathcal{B}$ to $\mathcal{G}$. Both of these effects are due to determinism, as can be seen by the fact that, although the recovery rate for $\mathcal{A} \longrightarrow \mathcal{G}$ is somewhat lower (69.4%) in the second series, the one for $\mathcal{B} \longrightarrow \mathcal{G}$ jumps to 61.6%. Also, the false orientation of the edge between $\mathcal{B}$ and $\mathcal{G}$ shrinks to 6%. Finally, there is one noticeable mistake, unrelated to constitutional inference, that is committed more frequently in the indeterministic setting: in series 1, PC issues an arrow from $\mathcal{Y}$ to $\mathcal{S}$ in 29.3% of the trials; that number increases to 40.9% in series 2.

Our second experiment differs from the first only insofar as Fisher's Z is replaced by the non-parametric independence test RCIT (Strobl, Zhang, and Visweswaran [2018]). We re-analyse the same 2000 data sets with a sample size of 10 000 each, using the same tuning parameters as before. In a first trial series (analysing the first 1000 data sets), $\mathcal{G}$ is a deterministic function of $\mathcal{A}$ and $\mathcal{B}$, in a second series, all variables are sampled with an error term. The purpose of the second experiment is to scrutinize PC's performance in analysing mixed variable sets when it is run with a non-parametric independence test, that is, a test not relying on partial correlations.

There exist various non-parametric independence tests, but most of them suffer from severe performance limitations and cannot process sample sizes of 10 000 observations in reasonable time (on an ordinary computer). We chose RCIT because of its computational efficiency, which allowed us to run the same tests as in the first experiment, and because of its promising benchmarking track record (see Strobl, Zhang, and Visweswaran [2018]).[10]

The results of the second experiment are plotted in Figure 7. The first thing to note is that, when PC is run with RCIT, it recognizes the independencies induced by determinism much more reliably than with Fisher's Z. In series 1, PC outputs $\mathcal{Y} \longrightarrow \mathcal{G}$ in only 27% of the trials, $\mathcal{I} \longrightarrow \mathcal{G}$ in 7.8%, and $\mathcal{G} \longrightarrow \mathcal{S}$ in 25.6%. Though these edges appear even less (17.3%, 5.1%, and 14.5%, respectively) in series 2, they are issued more frequently than in series 2 of the first experiment, where the indeterministic data is processed with Fisher's Z.

Next, in both trial series of the second experiment, the completeness ratios for edges and orientations are much lower than in the first experiment. When run with RCIT, PC, on average, recovers only 51.1% of the true edges and only 28.5% of the true orientations under determinism. Correspondingly, the constitutive arrow $\mathcal{A} \longrightarrow \mathcal{G}$ is recovered merely half as much (37.9%) in series 1 of the second experiment as it is in series 1 of the first experiment. Under indeterminism, the completeness ratios and the recovery rates of the arrows connecting $\mathcal{A}$ and $\mathcal{B}$ to $\mathcal{G}$ improve a bit but they do not reach the corresponding

[10] While, for example, **bnlearn** (Scutari [2010]) provides ready-made functionalities for running PC with various non-parametric independence tests, RCIT is not yet available in standard PC software. We are grateful to Eric Strobl for sharing a wrapper function of RCIT with us that can be called from within the `pc()` function of the **pcalg** package.

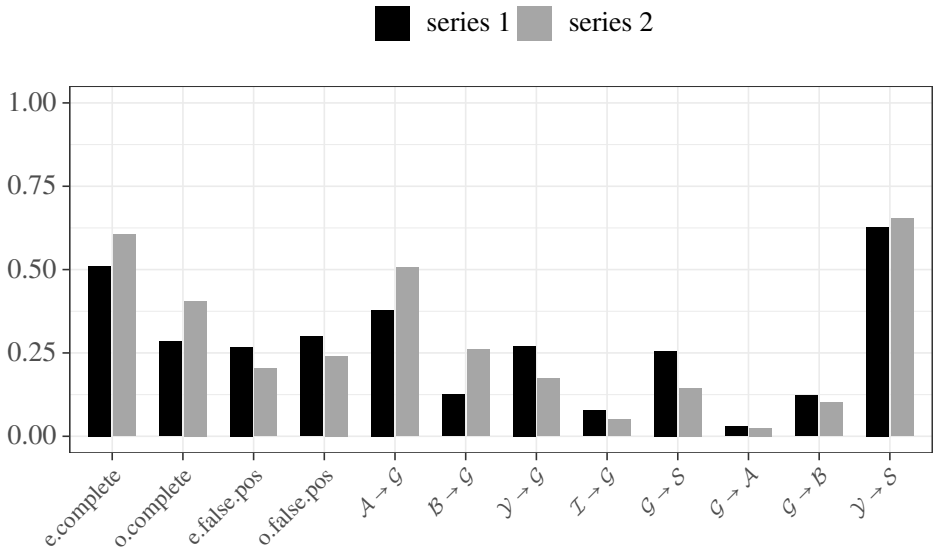

Figure 7: Results of the second benchmark experiment running PC with the non-parametric independence test RCIT.

rates of the first experiment. More concretely, even though series 2 features ideal discovery conditions for PC, PC only finds 40.6% of the true arrows when it is run with RCIT. That performance is clearly unsatisfactory.

Yet, despite the fact that PC with RCIT infers much less adjacencies overall it does not commit less false positives. When $\mathcal{G}$ is a deterministic function of $\mathcal{A}$ and $\mathcal{B}$, the false positive ratios for both edges (26.8%) and orientations (30.0%) are about the same as when PC is run with Fisher's Z. But when $\mathcal{G}$ is sampled with an error, PC issues twice as many false positives with RCIT: 20.3% false adjacencies and 24.1% false orientations. In sum, even though PC with RCIT recognizes the independencies induced by determinism, its inferences from data comprising both causally and constitutively related variables are not more reliable than when PC is run with Fisher's Z. On the one hand, PC with RCIT makes more mistakes of a different sort; for example, it issues an arrow $\mathcal{Y} \longrightarrow \mathcal{S}$ in 62.8% of the trials of series 1 (which number even increases in series 2). On the other hand, since PC infers less adjacencies with RCIT overall, mistakes weigh more heavily on the averaged false positive scores.

Correctly analysing systems featuring constitutively and causally related variables by means of PC, either with parametric or non-parametric independence tests, is an intricate and error-prone matter, even when the system is linear, with plenty of unshielded colliders, and the sample size is large. In the deterministic trial series of both experiments, the probability that PC identifies $\mathcal{B}$ as a constituent of $\mathcal{G}$ is only half as high as the probability of inferring a false orientation. This clearly suggests a negative answer to the question whether PC could reliably infer the presence of causal/constitutive dependence relations in mechanistic systems. In our (paradigmatic) test structure, the prospect of correctly identifying the

constituents of $\mathcal{G}$ is way too low to counterbalance the risk of a false positive.

8 Conclusion

Alexander Gebharter has recently claimed that the PC algorithm may be fruitfully brought to bear on the task of constitutive discovery. He proposes that it be used to infer causal as well as constitutive dependencies in one go, despite the widespread view that causation and constitution are fundamentally different kinds of dependence relations.

In this paper, we argued that Gebharter severely underestimates the problems constitutive relations and the features of his discovery approach induce for PC. First, the two-level restriction, which Gebharter introduces to warrant CFC, renders the justification of CMC more problematic than in ordinary causal discovery contexts. Second, CFC is systematically violated even in mechanistic systems complying to the two-level restriction, meaning that PC cannot be reliably applied. Third, the latter problem cannot be remedied by employing a modified version of PC, namely PCD, that is designed for contexts of CFC violations induced by determinism. The reason is that constitutive dependencies tend to generate probabilistic independencies that are unfaithful even by PCD's weakened faithfulness standards. Fourth, only interpreting the presence (and not the absence) of edges in outputs of the PC algorithm produced in CFC-violating contexts does not amount to a promising weakening of Gebharter's proposal. We showed, in two extended benchmarking experiments, that determinism induced by constitution prevents PC from reliably inferring the presence of causal/constitutive dependencies.

From all this, we conclude that Gebharter's proposal to use PC for constitutive discovery is a nonstarter. PC is an algorithm custom-built for causal discovery contexts, which are characterized by (pseudo)indeterministic dependencies. Deterministic dependencies as induced by constitution are beyond PC's scope. Ultimately, this finding casts doubt on Gebharter's starting point, *viz.* the assumption that constitution may be treated as a form of deterministic direct causation complying with CMC and CFC.

Appendix

R Replication Script for Experiment One

```
# Required R packages
library(pcalg)
library(ggplot2)
library(reshape)
library(tikzDevice)

# Auxiliary Functions
sortString <- function(x){
  if(length(x)>0) {
  r <- strsplit(x, "\\|")
  r <- lapply(r, sort)
  lapply(r, paste0, collapse = "|")} else {
  x
  }
}
```

```
bidir <- function(x)  {
  if(length(x)>0){
  r <- expand.grid(x,x)
  r <- r[apply(r, 1, function(x) all(!duplicated(x))),]
  resul <- vector("logical",nrow(r))
  for(i in 1:nrow(r)){
    resul[i] <-  identical(sortString(as.character(r[i,1])),sortString(as.character(r[i,2])))
  }
  r <- cbind(r,resul)
  r <- as.matrix(r[which(r[,3]==TRUE),])
  unique(as.vector(r[,1:2]))
  }
}

# Define test variants
variants <- c("series␣1", "series␣2")

# Define true graph
true.graph <- c("I|A", "I|B", "I|S", "A|B", "A|S", "Y|B", "A|G", "B|S", "B|G", "Z|A", "X|S")

# Number of trials
n <- 1000
# Sample Size
sampleSize <- 10000

# Draw seeds for replication
suppressWarnings(RNGversion("3.6.0")) # enforcing sampling method from R 3.6
set.seed(48)
seeds.param <- sample(.Machine$integer.max, n)

# Initialize score lists
compare.variants <- score.list.storage <- data.storage <- analyses.storage <- vector("list", length(variants))

# Variant Loop
# ------------

for(m in 1:length(variants)){
  cat(m,  "variant","\n")

  variant <- variants[m]

  # Repetition Loop
  # -------------
  score.list.rep <- edges.list.rep <- data.list.rep <- analyses.rep <- vector("list", n)

  for(k in 1:n){
  cat(k, "\n")
  set.seed(seeds.param[k])

     # Sample error terms
     errorA <- rnorm(sampleSize,0)
     errorB <- rnorm(sampleSize,0)
     errorS <- rnorm(sampleSize,0)
     errorG <- rnorm(sampleSize,0)

     # Sample exogenous variables
     p.range <- setdiff(-5:5,0)  # parameter range
     I <- sample(p.range,1) + rnorm(sampleSize,0,sample(1:10,1))
     X <- sample(p.range,1) + rnorm(sampleSize,0,sample(1:10,1))
     Z <- sample(p.range,1) + rnorm(sampleSize,0,sample(1:10,1))
     Y <- sample(p.range,1) + rnorm(sampleSize,0,sample(1:10,1))

     # Sample endogenous parts A and B
     A <- sample(p.range,1)*I + sample(p.range,1)*Z + errorA
     B <- sample(p.range,1)*I + sample(p.range,1)*A + sample(p.range,1)*Y + errorB

     # Sample G
     if(variant=="series␣1"){G <- sample(p.range,1)*A + sample(p.range,1)*B
     } else {G <- sample(p.range,1)*A + sample(p.range,1)*B  + errorG
     }

     # Sample S
     S <- sample(p.range,1)*I + sample(p.range,1)*A + sample(p.range,1)*B +
                                             sample(p.range,1)*X + errorS

     # Generate data list
     data <- cbind(G,A,B,X,Z,Y,I,S)

     # Store data
     data.list.rep[[k]] <- data

     # Run PC
     V <- colnames(data)
     suffStat <- list(C = cor(data), n = nrow(data))
     analysis <- pc(suffStat, indepTest = gaussCItest, alpha = 0.05, labels = V, maj.rule = T)
     analyses.rep[[k]] <- analysis

     # Build all edges in graph
     edgesGraph <- arrowsGraph <-  gsub(".weight",'',
                                        attr(unlist(analysis@graph@edgeData@data),"names"))
     edgesGraph <- unlist(sortString(edgesGraph))

     # Build all possible (un)directed edges
     allEdges <- combn(colnames(data),2)
     allEdges <- apply(allEdges, 2, function(x) paste0(x[1],"|",x[2]))
     allEdges <- unlist(sortString(allEdges))

     allArrows <- expand.grid(colnames(data),colnames(data))
     allArrows <- allArrows[apply(allArrows, 1, function(x) all(!duplicated(x))),]
```

```
    allArrows <- as.vector(apply(as.matrix(allArrows), 1, function(x) paste0(x[1],"|",x[2])))

    # Score undirected edges recovered
    score <- allEdges %in% edgesGraph
    score <- as.data.frame(matrix(score,nrow=1,ncol=length(allEdges),
           byrow = T,dimnames = list(1,allEdges)))

    # False positives in undirected edges/AdjCompletness (true edges in Graph/edges in true graph)
    x <- unique(unlist(sortString(true.graph)))
    Adj.false.pos <- if(length(edgesGraph)>0){
                      length(setdiff(unique(unlist(sortString(edgesGraph))),x))/
                                                  length(unique(edgesGraph))
                      }else{0}
    AdjCompleteness <- length(intersect(unique(unlist(sortString(edgesGraph))),x))/length(x)
    score$Adj.false.pos <- Adj.false.pos
    score$AdjCompleteness <- AdjCompleteness

    # False positives in arrows/ArrCompleteness (Arr.true.pos/directed arrows in true graph)
    x <- true.graph
    z <- bidir(arrowsGraph) # bi- and undirectional edges in Graph
    dirArr <- setdiff(arrowsGraph,z) # directed edges in Graph
    Arr.false.pos <- if(length(dirArr)>0){
                      length(union(setdiff(dirArr,x),
                                   setdiff(unlist(sortString(z)), unlist(sortString(x)))))/
                                                          length(unique(edgesGraph))
                      }else{0}
    ArrCompleteness <- length(intersect(dirArr,x))/length(x)

    score$Arr.false.pos <- Arr.false.pos
    score$ArrCompleteness <- ArrCompleteness
    score$pathAG <- "A|G" %in% dirArr
    score$pathBG <- "B|G" %in% dirArr
    score$pathYS <- "Y|S" %in% dirArr
    score$pathYG <- "Y|G" %in% dirArr
    score$pathZG <- "Z|G" %in% dirArr
    score$pathIG <- "I|G" %in% dirArr
    score$pathGS <- "G|S" %in% dirArr
    score$pathGA <- "G|A" %in% dirArr
    score$pathGB <- "G|B" %in% dirArr
    score.list.rep[[k]] <-  score

    } # End repetition loop

   # Overall scoring
   overall.score <- do.call(rbind,score.list.rep)
   overall.ratio <-  matrix(1, ncol = ncol(overall.score), byrow = T,
           dimnames = list(1,names(overall.score)))

   for(i in 1:length(allEdges)){
       overall.ratio[1,i] <- length(which(overall.score[,i]))/nrow(overall.score)
   }
       overall.ratio[,c("Adj.false.pos")] <- mean(overall.score$Adj.false.pos)
       overall.ratio[,c("Arr.false.pos")] <- mean(overall.score$Arr.false.pos)
       overall.ratio[,c("AdjCompleteness")] <- mean(overall.score$AdjCompleteness)
       overall.ratio[,c("ArrCompleteness")] <- mean(overall.score$ArrCompleteness)
       overall.ratio[,c("pathAG")] <- length(which(overall.score[,c("pathAG")]))/nrow(overall.score)
       overall.ratio[,c("pathBG")] <- length(which(overall.score[,c("pathBG")]))/nrow(overall.score)
       overall.ratio[,c("pathYS")] <- length(which(overall.score[,c("pathYS")]))/nrow(overall.score)
       overall.ratio[,c("pathYG")] <- length(which(overall.score[,c("pathYG")]))/nrow(overall.score)
       overall.ratio[,c("pathZG")] <- length(which(overall.score[,c("pathZG")]))/nrow(overall.score)
       overall.ratio[,c("pathIG")] <- length(which(overall.score[,c("pathIG")]))/nrow(overall.score)
       overall.ratio[,c("pathGS")] <- length(which(overall.score[,c("pathGS")]))/nrow(overall.score)
       overall.ratio[,c("pathGA")] <- length(which(overall.score[,c("pathGA")]))/nrow(overall.score)
       overall.ratio[,c("pathGB")] <- length(which(overall.score[,c("pathGB")]))/nrow(overall.score)

    # Store analyses, data, scores
    analyses.storage[[m]] <- analyses.rep
    data.storage[[m]] <-  data.list.rep
    score.list.storage[[m]] <- score.list.rep
    compare.variants[[m]] <- overall.ratio

}  # END variants loop

# Average number of edges per trial
# ---------------------------------
# Series 1
ana_sum_series1 <- lapply(score.list.storage[[1]], function(x) sum(as.matrix(x[1:28])))
mean(unlist(ana_sum_series1))

# Series 2
ana_sum_series2 <- lapply(score.list.storage[[2]], function(x) sum(as.matrix(x[1:28])))
mean(unlist(ana_sum_series2))

# Plot 1 (Figure 6)
# -----------------
selection <- lapply(compare.variants, function(x) x[,c("AdjCompleteness","ArrCompleteness",
                                      "Adj.false.pos","Arr.false.pos", "pathAG", "pathBG",
                                      "pathYG",  "pathIG", "pathGS","pathGA", "pathGB",
                                      "pathYS")])
selection <- lapply(selection, setNames, c("edge.complete","orient.complete","edge.false.pos",
                                      "orient.false.pos",
                                      "A->G", "B->G", "Y->G", "I->G", "G->S", "G->A",
                                      "G->B", "Y->S"))
final.score <-  as.data.frame(do.call(rbind,selection))
final.score$variants <- factor(variants, levels = variants)
k1 <- melt(final.score, id.vars = "variants")
colnames(k1) <- c("variants", "property", "value")
plot1 <- ggplot(k1, aes(x = property, y = value,group = 1, fill = variants)) +
    geom_bar(stat = "identity", position = position_dodge2(), width = .7) +
```

```
    scale_fill_grey(start = 0, end = .65) +
    theme_bw() + ylim(0, 1) + theme(legend.position = "top") +
    theme(legend.title = element_blank()) +
    theme(plot.title = element_text(size = 9))+ theme(axis.text.x = element_text(size = 8,
                                                      angle = 45,hjust = 1)) +
    scale_x_discrete(name ="")

options(tz="CA")
tikz(file = "plot1.tex", width = 5.1, height = 3.4)
print(plot1)
dev.off()
getwd()


# Running PC with Fisher's Z on a shrunk correlation matrix (cf. footnote 9)
# --------------------------------------------------------------------------
# Use the data set from trial 234 of series 1
# (can be changed to any trial number between 1 and 1000)
data <- data.storage[[1]][[234]]
V <- colnames(data)
suffStat <- list(C = round(corpcor::pcor.shrink(data, lambda = 1e-10), 3), n = nrow(data))
analysis <- pc(suffStat, indepTest = gaussCItest, alpha = 0.05, labels = V, maj.rule = T)
plot(analysis)
```

R Replication Script for Experiment Two

```
# Required R packages
library(pcalg)
library(ggplot2)
library(reshape)
library(tikzDevice)
library(RCIT)

# Auxiliary Functions
sortString <- function(x){
  if(length(x)>0) {
  r <- strsplit(x, "\\|")
  r <- lapply(r, sort)
  lapply(r, paste0, collapse = "|")} else {
  x
  }
}

bidir <- function(x)  {
  if(length(x)>0){
  r <- expand.grid(x,x)
  r <- r[apply(r, 1, function(x) all(!duplicated(x))),]
  resul <- vector("logical",nrow(r))
  for(i in 1:nrow(r)){
    resul[i] <-  identical(sortString(as.character(r[i,1])),sortString(as.character(r[i,2])))
  }
  r <- cbind(r,resul)
  r <- as.matrix(r[which(r[,3]==TRUE),])
  unique(as.vector(r[,1:2]))
  }
}

# RCTI wrapper function, thanks to Eric Strobl for sharing
RCIT_wrap <- function(x, y, z, suffStat){
  x1=suffStat$data[,x];
  y1=suffStat$data[,y];
  z1=suffStat$data[,z];
  out = RCIT:::RCIT(x1, y1, z1, num_f = 25);
  return(out$p)
}

# Define test variants
variants <- c("series␣1", "series␣2")

# Define true graph
true.graph <- c("I|A", "I|B", "I|S", "A|B", "A|S", "Y|B", "A|G", "B|S", "B|G", "Z|A", "X|S")

# Number of trials
n <- 1000
# Sample Size
sampleSize <- 10000

# Draw seeds for replication
suppressWarnings(RNGversion("3.6.0")) # enforcing sampling method from R 3.6
set.seed(48)
seeds.param <- sample(.Machine$integer.max, n)

# Initialize score lists
compare.variants <- score.list.storage <- data.storage <- analyses.storage <- vector("list", length(variants))

# Variant Loop
# ------------

for(m in 1:length(variants)){
  cat(m,  "variant","\n")

  variant <- variants[m]

  # Repetition Loop
```

```
# -------------
score.list.rep <- edges.list.rep <- data.list.rep <- analyses.rep <- vector("list", n)

for(k in 1:n){
cat(k, "\n")
set.seed(seeds.param[k])

   # Sample error terms
   errorA <- rnorm(sampleSize,0)
   errorB <- rnorm(sampleSize,0)
   errorS <- rnorm(sampleSize,0)
   errorG <- rnorm(sampleSize,0)
   # errorC <- rnorm(sampleSize,0)

   # Sample exogenous variables
   p.range <- setdiff(-5:5,0) # parameter range
   I <- sample(p.range,1) + rnorm(sampleSize,0,sample(1:10,1))
   X <- sample(p.range,1) + rnorm(sampleSize,0,sample(1:10,1))
   Z <- sample(p.range,1) + rnorm(sampleSize,0,sample(1:10,1))
   Y <- sample(p.range,1) + rnorm(sampleSize,0,sample(1:10,1))

   # Sample endogenous parts A and B
   A <- sample(p.range,1)*I + sample(p.range,1)*Z + errorA
   B <- sample(p.range,1)*I + sample(p.range,1)*A + sample(p.range,1)*Y + errorB

   # Sample G
   if(variant=="series 1"){G <- sample(p.range,1)*A + sample(p.range,1)*B
   } else {G <- sample(p.range,1)*A + sample(p.range,1)*B  + errorG
   }

   # Sample S
   S <- sample(p.range,1)*I + sample(p.range,1)*A + sample(p.range,1)*B +
                                          sample(p.range,1)*X + errorS

   # Generate data list
   data <- cbind(G,A,B,X,Z,Y,I,S)

   # Store data
   data.list.rep[[k]] <- data

   # Run PC
   V <- colnames(data)
   suffStat = list(data = data)
   analysis <-  pc(suffStat, indepTest = RCIT_wrap, alpha = 0.05, labels = V, maj.rule = T)
   analyses.rep[[k]] <- analysis

   # Build all edges in graph
   edgesGraph <- arrowsGraph <-  gsub(".weight",'',
                                  attr(unlist(analysis@graph@edgeData@data),"names"))
   edgesGraph <- unlist(sortString(edgesGraph))

   # Build all possible (un)directed edges
   allEdges <- combn(colnames(data),2)
   allEdges <- apply(allEdges, 2, function(x) paste0(x[1],"|",x[2]))
   allEdges <- unlist(sortString(allEdges))

   allArrows <- expand.grid(colnames(data),colnames(data))
   allArrows <- allArrows[apply(allArrows, 1, function(x) all(!duplicated(x))),]
   allArrows <- as.vector(apply(as.matrix(allArrows), 1, function(x) paste0(x[1],"|",x[2])))

   # Score undirected edges recovered
   score <- allEdges %in% edgesGraph
   score <- as.data.frame(matrix(score,nrow=1,ncol=length(allEdges),
          byrow = T,dimnames = list(1,allEdges)))

   # False positives in undirected edges/AdjCompletness (true edges in Graph/edges in true graph)
   x <- unique(unlist(sortString(true.graph)))
   Adj.false.pos <- if(length(edgesGraph)>0){
                       length(setdiff(unique(unlist(sortString(edgesGraph))), x))/
                                                    length(unique(edgesGraph))
                     }else{0}
   AdjCompleteness <- length(intersect(unique(unlist(sortString(edgesGraph))),x))/length(x)
   score$Adj.false.pos <- Adj.false.pos
   score$AdjCompleteness <- AdjCompleteness

   # False positives in arrows/ArrCompleteness (Arr.true.pos/directed arrows in true graph)
   x <- true.graph
   z <- bidir(arrowsGraph) # bi- and undirectional edges in Graph
   dirArr <- setdiff(arrowsGraph,z) # directed edges in Graph
   Arr.false.pos <- if(length(dirArr)>0){
                       length(union(setdiff(dirArr,x),
                                    setdiff(unlist(sortString(z)), unlist(sortString(x))))/
                                                       length(unique(edgesGraph))
                       }else{0}
   ArrCompleteness <- length(intersect(dirArr,x))/length(x)

   score$Arr.false.pos <- Arr.false.pos
   score$ArrCompleteness <- ArrCompleteness
   score$pathAG <- "A|G" %in% dirArr
   score$pathBG <- "B|G" %in% dirArr
   score$pathYS <- "Y|S" %in% dirArr
   score$pathYG <- "Y|G" %in% dirArr
   score$pathZG <- "Z|G" %in% dirArr
   score$pathIG <- "I|G" %in% dirArr
   score$pathGS <- "G|S" %in% dirArr
   score$pathGA <- "G|A" %in% dirArr
   score$pathGB <- "G|B" %in% dirArr
   score.list.rep[[k]] <-  score
```

```
    } # End repetition loop

  # Overall scoring
  overall.score <- do.call(rbind,score.list.rep)
  overall.ratio <-  matrix(1, ncol = ncol(overall.score), byrow = T,
          dimnames = list(1,names(overall.score)))


  for(i in 1:length(allEdges)){
      overall.ratio[1,i] <- length(which(overall.score[,i]))/nrow(overall.score)
  }
      overall.ratio[,c("Adj.false.pos")] <- mean(overall.score$Adj.false.pos)
      overall.ratio[,c("Arr.false.pos")] <- mean(overall.score$Arr.false.pos)
      overall.ratio[,c("AdjCompleteness")] <- mean(overall.score$AdjCompleteness)
      overall.ratio[,c("ArrCompleteness")] <- mean(overall.score$ArrCompleteness)
      overall.ratio[,c("pathAG")] <- length(which(overall.score[,c("pathAG")]))/nrow(overall.score)
      overall.ratio[,c("pathBG")] <- length(which(overall.score[,c("pathBG")]))/nrow(overall.score)
      overall.ratio[,c("pathYS")] <- length(which(overall.score[,c("pathYS")]))/nrow(overall.score)
      overall.ratio[,c("pathYG")] <- length(which(overall.score[,c("pathYG")]))/nrow(overall.score)
      overall.ratio[,c("pathZG")] <- length(which(overall.score[,c("pathZG")]))/nrow(overall.score)
      overall.ratio[,c("pathIG")] <- length(which(overall.score[,c("pathIG")]))/nrow(overall.score)
      overall.ratio[,c("pathGS")] <- length(which(overall.score[,c("pathGS")]))/nrow(overall.score)
      overall.ratio[,c("pathGA")] <- length(which(overall.score[,c("pathGA")]))/nrow(overall.score)
      overall.ratio[,c("pathGB")] <- length(which(overall.score[,c("pathGB")]))/nrow(overall.score)

   # Store analyses, data, scores
   analyses.storage[[m]] <- analyses.rep
   data.storage[[m]] <-  data.list.rep
   score.list.storage[[m]] <- score.list.rep
   compare.variants[[m]] <- overall.ratio

}  # END variants loop

# Average number of edges per trial
# ---------------------------------
# Series 1
ana_sum_series1 <- lapply(score.list.storage[[1]], function(x) sum(as.matrix(x[1:28])))
mean(unlist(ana_sum_series1))

# Series 2
ana_sum_series2 <- lapply(score.list.storage[[2]], function(x) sum(as.matrix(x[1:28])))
mean(unlist(ana_sum_series2))

# Plot 2 (Figure 7)
# -----------------
selection <- lapply(compare.variants, function(x) x[,c("AdjCompleteness", "ArrCompleteness",
                                                       "Adj.false.pos", "Arr.false.pos", "pathAG",
                                                       "pathBG", "pathYG", "pathIG", "pathGS",
                                                       "pathGA", "pathGB", "pathYS")])
selection <- lapply(selection, setNames, c("edge.complete", "orient.complete", "edge.false.pos",
                                           "orient.false.pos",
                                            "A->G", "B->G", "Y->G", "I->G", "G->S", "G->A",
                                            "G->B", "Y->S"))
final.score <-  as.data.frame(do.call(rbind, selection))
final.score$variants <- factor(variants, levels = variants)
k1 <- melt(final.score, id.vars = "variants")
colnames(k1) <- c("variants", "property", "value")
plot2 <- ggplot(k1, aes(x = property, y = value,group = 1, fill = variants)) +
  geom_bar(stat = "identity", position = position_dodge2(), width = .7) +
  scale_fill_grey(start = 0, end = .65)+
  theme_bw() + ylim(0, 1) + theme(legend.position="top") +
  theme(legend.title = element_blank()) +
  theme(plot.title = element_text(size = 9)) +
  theme(axis.text.x = element_text(size = 8, angle = 45, hjust = 1)) +
  scale_x_discrete(name = "")

options(tz="CA")
tikz(file = "plot2.tex", width = 5.1, height = 3.4)
print(plot2)
dev.off()
getwd()
```


Acknowledgements

We thank the audiences of Explanatory Power, Geneva, 15 June 2018, and Causation, Mechanism and Difference-Makers, Copenhagen, 3 August 2018. We are especially grateful to Alexander Gebharter, Beate Krickel, Daniel Malinsky, Alessio Moneta, Joseph Ramsey, Marco Scutari, and Kun Zhang for helpful discussions, and to three anonymous reviewers for comments on earlier versions of the paper. We also thank Eric Strobl for sharing a wrapper function, which allowed us to run PC with the non-parametric test RCIT.

Funding

This research was generously supported by the Swiss National Science Foundation (grants no. CRSII 1_147685/1 and 100012E_160866/1 for LC and grant no. PP00P1_144736/1 for MB) and the Bergen Research Foundation (grant no. 811886 for MB).

Lorenzo Casini
University of Geneva
Department of Philosophy
Geneva, Switzerland
lorenzo.casini@unige.ch

Michael Baumgartner
University of Bergen
Department of Philosophy
Bergen, Norway
michael.baumgartner@uib.no

References

Baumgartner, M. and L. Casini [2017]: 'An abductive theory of constitution', *Philosophy of Science*, **84**, pp. 214–33.

Bechtel, W. and A. Abrahamsen [2005]: 'Explanation: a mechanist alternative', *Studies in the History and Philosophy of the Biological and Biomedical Sciences*, **36**, pp. 421–41.

Colombo, D. and M. H. Maathuis [2014]: 'Order-independent constraint-based causal structure learning', *Journal of Machine Learning Research*, **15**, pp. 3921–62.

Couch, M. B. [2011]: 'Mechanisms and constitutive relevance', *Synthese*, **183**, pp. 375–88.

Craver, C. F. [2007]: *Explaining the brain*, Oxford: Oxford University Press.

Edwards, D. I. [2000]: *Introduction to graphical modelling* (second ed.), New York: Springer.

Eronen, M. I. [2011]: *Reduction in philosophy of mind: A pluralistic account*, Frankfurt am Main: Ontos.

Eronen, M. I. [2013]: 'No levels, no problems: Downward causation in neuroscience', *Philosophy of Science*, **80**, pp. 1042–52.

Gebharter, A. [2017a]: 'Causal exclusion and causal Bayes nets', *Philosophy and Phenomenological Research*, **95**, pp. 353–75.

Gebharter, A. [2017b]: 'Uncovering constitutive relevance relations in mechanisms', *Philosophical Studies*, **174**, pp. 2645–66.

Gentle, J. E. [2013]: *Theory of Statistics*, George Mason University.

Glennan, S. [1996]: 'Mechanisms and the nature of causation', *Erkenntnis*, **44**, pp. 49–71.

Glennan, S. [2002]: 'Rethinking mechanistic explanation', *Philosophy of Science*, **69**, pp. 342–53.

Glymour, C. [2007]: 'Learning the structure of deterministic systems', in A. Gopnik and L. Schulz (eds), *Causal learning: psychology, philosophy, and computation*, Oxford: Oxford University Press, pp. 231–40.

Harbecke, J. [2010]: 'Mechanistic constitution in neurobiological explanations', *International Studies in the Philosophy of Science*, **24**, pp. 267–85.

Kalisch, M., M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann [2012]: 'Causal inference using graphical models with the R package pcalg', *Journal of Statistical Software*, **47**, pp. 1–26.

Machamer, P., L. Darden, and C. Craver [2000]: 'Thinking about mechanisms', *Philosophy of Science*, **67**, pp. 1–25.

Pearl, J. [2009]: *Causality: models, reasoning, and inference* (second ed.), Cambridge: Cambridge University Press.

Schäfer, J. and K. Strimmer [2005]: 'A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics', *Statistical Applications in Genetics and Molecular Biology*, **4**, doi: 10.2202/1544-6115.1175.

Scutari, M. [2010]: 'Learning bayesian networks with the bnlearn R package', *Journal of Statistical Software*, **35**, pp. 1–22.

Spirtes, P., C. Glymour, and R. Scheines [2000]: *Causation, prediction, and search* (second ed.), Cambridge, MA: MIT Press.

Spohn, W. [2006]: 'Causation: An alternative', *The British Journal for the Philosophy of Science*, **57**, pp. 93–119.

Strobl, E., K. Zhang, and S. Visweswaran [2018]: 'Approximate kernel-based conditional independence tests for fast non-parametric causal discovery', *Journal of Causal Inference*, doi: 10.1515/jci-2018-0017.

Wimsatt, W. [2007]: *Re-engineering philosophy for limited beings*, Cambridge, MA: Harvard University Press.

Zhalama, J. Zhang, and W. Mayer [2017]: 'Weakening faithfulness: some heuristic causal discovery algorithms', *International Journal of Data Science and Analytics*, **3**, pp. 93–104.

Zhang, J. [2006]: *Causal inference and reasoning in causally insufficient systems*, Ph. D. thesis, Department of Philosophy, Carnegie Mellon University.

Zhang, J. and P. Spirtes [2008]: 'Detection of unfaithfulness and robust causal inference', *Minds and Machines*, **18**, pp. 239–71.

Zhang, J. and P. Spirtes [2016]: 'The three faces of faithfulness', *Synthese*, **193**, pp. 1011–27.