

Faithfulness, Coordination and Causal Coincidences

Naftali Weinberger¹

Received: 2 March 2016 / Accepted: 21 January 2017

© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract Within the causal modeling literature, debates about the Causal Faithfulness Condition (CFC) have concerned whether it is probable that the parameters in causal models will have values such that distinct causal paths will cancel. As the parameters in a model are fixed by the probability distribution over its variables, it is initially puzzling what it means to assign probabilities to these parameters. I propose that to assign a probability to a parameter in a model is to treat that parameter as a function of a variable in an augmented model. By combining this proposal with widely adopted principles regarding which variables must be included in a model, I argue that the various proposed counterexamples to CFC involving coordinated parameters are not genuine counterexamples. I then consider the cases in which CFC fails due not to coordination, but by coincidence, and propose explanatory and predictive bases for ruling out such coincidences without presuming that they are improbable. The aim of the proposed defenses is not to show that CFC never fails, but rather to argue that its use in a particular context may be defended using general modeling assumptions rather than by relying on claims about how often it fails.

1 Introduction

Graphical causal models enable one to represent causal hypotheses in terms of the functional dependence relations among a set of random variables. Given assumptions about the relationship between causal hypotheses and probability distributions, one can use the joint probability distribution over the variables to narrow down the set of hypotheses considered. One common assumption is the Causal Faithfulness Condition

✉ Naftali Weinberger
Naftali.weinberger@gmail.com

¹ Tilburg Center for Logic, Ethics and Philosophy of Science, Tilburg University,
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

(CFC, or ‘faithfulness’), which says that variables that are causally connected in a particular way in the graph are probabilistically dependent. One way for CFC to fail is if a variable is causally related to its effect along two paths that cancel out, thereby rendering the cause and effect probabilistically independent. Despite this possibility, CFC is often defended on the grounds that it is highly improbable that two causal paths will cancel out exactly (Spirtes et al. 2000). In response, Cartwright (1999) and Hoover (2001) argue that in cases where a policymaker seeks to make the two paths cancel, failures of faithfulness *are* probable. Additionally, Andersen (2013) contends that cancellation is probable in homeostatic biological systems in which a quantity is maintained at its equilibrium value.

While both proponents of CFC and those who note its limitations make assertions about the probability of cancellation, it is initially puzzling what the basis for such probability attributions could be. In causal modeling, one uses a joint probability distribution over the variables in the model to estimate the parameters that represent how an effect depends on its causes. To talk about the probability of cancellation is not to talk about a feature of the joint probability distribution, but about the probability that the parameters estimated from this distribution will take on certain values (such that they cancel). As these parameters are fixed by the joint distribution over the variables, what is the basis for assigning them their own distribution? Here I propose that to assign a probability to a parameter in a causal model is to treat that parameter as a function of a variable in an augmented model. Explicitly considering these augmented models yields insights about the conditions under which cancellation occurs. The counterexamples to faithfulness described above involve cases in which the paths do not cancel by chance, but as a result of a process that coordinates the model’s parameters. In the augmented models, this coordination results either from common causes or common effects of variables in the non-augmented models, and there are standard causal modeling assumptions that require one to include these additional variables. Once one does so, it becomes clear that certain apparent failures of faithfulness in non-augmented models are not failures of faithfulness in the distribution corresponding to augmented model. This, in fact, is what occurs in the standard counterexamples. More generally, there will not be a failure of faithfulness as long as there are values of the variables in the augmented model for which there is not cancellation.

In cases where a model’s parameters are not coordinated, there may still be cancellation, but this would be coincidental in the same sense that it would be coincidental for two uncorrelated variables to take on the same value. This may seem to support arguments showing that failures of faithfulness are improbable. Yet the claim that a cancellation is coincidental does not entail that is improbable. In cases where one does not have adequate information to determine whether cancellation is probable, I advocate predictive and explanatory grounds for preferring models that do not require cancellation to account for the probability distribution. In presenting these arguments, I do not aim to show that we should always adopt faithfulness and never worry about its failure. Rather, it is to argue that when thinking about whether faithfulness is justified in a particular domain, we do not necessarily need to determine whether it is likely to fail, but instead can appeal to more general modeling principles.

The failures of faithfulness discussed in this article (and in most of the philosophical literature) are only a subset of the possible failures of faithfulness. More specifically, the failures considered all violate a principle known as triangle-faithfulness, which is a consequence of CFC. Accordingly, it may turn out that the arguments I present are only relevant to one consequence of faithfulness. Nevertheless, triangle-faithfulness is entailed even by logically weaker alternatives to faithfulness such as (P-)minimality (Pearl 2009; Zhang 2013) and frugality (Forster et al. forthcoming) and failures of triangle-faithfulness pose special problems for causal inference (Zhang and Spirtes 2008). Triangle-faithfulness is thus important in its own right and merits an extended discussion of the conditions under which it is justified.

This paper proceeds as follows. Section 2 introduces the Causal Faithfulness Condition and distinguishes among different types of failures of faithfulness. Section 3 presents a debate regarding whether (triangle-) failures of faithfulness are probable and explains why the assignment of probabilities to parameters is puzzling. Section 4 presents the augmented variable strategy as a means of discussing probability distributions over parameters and uses this strategy to reveal that Cartwright and Hoover's counterexamples implicitly omit a common cause. Section 5 argues that non-coincidental failures will generally involve variables that are not part of the cancelling paths and that in the standard counterexamples faithfulness does not fail in the probability distribution including these additional variables. Section 6 presents two bases for preferring models that do not involve coincidental cancellation. Section 7 concludes.

2 The Causal Faithfulness Condition

It is increasingly common to use directed acyclic graphs (DAGs) to model the causal relationships among a set of variables. A DAG contains variables as nodes and directed edges (i.e. arrows) representing causal relationships between the nodes. X is a *parent* of Y when there is a causal arrow from X to Y . DAGs are associated with a set of equations in which each variable is a function of its parents and an unmeasured error term. These equations specify how intervening on a parent of a variable while holding the others fixed via interventions would change that variable's value.

Spirtes et al. (2000) and Pearl (2009) pioneered the use of DAGs in causal inference. Their frameworks enable one to systematically relate DAGs to probability distributions. When a DAG is *causally sufficient*—that is, all common causes of variables in the DAG are included in the DAG—two conditions suffice to create a mapping between a DAG and the probabilistic dependence and independence relations among the variables in a probability distribution: The *Causal Markov Condition* (CMC) and the *Causal Faithfulness Condition* (CFC, referred to as *stability* in Pearl¹). Given a DAG, CMC specifies which variables will

¹ Andersen (2013) claims that stability is distinct from faithfulness, though Pearl treats faithfulness and stability as synonymous and I know of no case in which the conditions come apart.

be probabilistically *independent* conditional on other variables and CFC specifies which ones will be probabilistically *dependent*. Even assuming CMC and CFC, the correct DAG will often be underdetermined by the probability distribution. That is, not every distinct DAG entails a distinct set of probabilistic dependencies.

To specify the sets of variables that are probabilistically independent according to CMC, it helps to introduce the graphical notion of *d-separation*. A path is *d-separated* by variable set \mathbf{Z} just in case²:

1. The path contains a triple $i \rightarrow m \rightarrow j$ or $i \leftarrow m \rightarrow j$ such that m is in \mathbf{Z}
2. The path contains a collider $i \rightarrow m \leftarrow j$ such that m is not in \mathbf{Z} and no descendant of m is in \mathbf{Z} . (Y is a *descendant* of X iff there is a path with unidirectional arrows from X to Y).

d-separation is property of paths (sets of connected arrows). Two variables are *d-separated* by \mathbf{Z} iff they are *d-separated* by \mathbf{Z} along all paths. The Causal Markov Condition states that all variables that are *d-separated* in a DAG will be conditionally independent in the corresponding probability distribution. The Causal Faithfulness Condition states that *only* the variables that are *d-separated* in a DAG will be independent (i.e. all others will be dependent). The importance of colliders is that conditioning on a collider will sometimes render its causes to be dependent. This is called endogenous selection bias (Elwert and Winship 2014). In contrast, conditioning on non-colliders along a path renders the variables linked by that path to be probabilistically independent (provided they are not *d-connected* along another path).

CMC is motivated by the idea that outside of cases of conditioning on a collider any probabilistic dependency between variables must be due to one variable causing the other or to their sharing a common cause. Whether it is the case that all such dependencies have causal explanations is controversial, though it is certainly *not* the case that all variable sets satisfy CMC. If a variable set V omits common causes of variables in V , then there can be dependencies among variables in V that are not causally accounted for by the variables in the set. For this reason, it is common when assuming CMC to also assume that one's variable set is causally sufficient. For similar reasoning, it is common to assume that one's variable set does not omit any common effects that have been conditioned on. Although CMC must be applied carefully, and may lead to unreliable inferences when applied to limited variable sets, the assumption that variable sets must be embeddable into broader variable sets that satisfy CMC is fundamental for causal inference. Without a condition such as CMC, it is not possible to make any causal inferences based on a probability distribution, since one would have no basis for specifying which distributions are compatible with a causal hypothesis.

While CMC may be justified via appeal to general principles about the relationship between causal hypotheses and probability distributions, CFC does not appear to be like this. It is not difficult to imagine counterexamples to CFC. It fails, for example, whenever there are two paths with equal and opposite effects between

² This definition is a paraphrase of Pearl's (2000, 16–17).

variables. Consider Fig. 1, which presents a model with linear parameters. If these parameters satisfy the constraint $\alpha\beta = -\gamma$, then X and Y will be uncorrelated, even though they are causally related. In what follows, I will call X the *treatment*, Y the *outcome* and Z the *mediator*.

Although the philosophical literature has focused on failures of faithfulness involving cancellation, these are not the only types of failures. For example, faithfulness fails in systems with deterministic relationships among variables. If A deterministically causes B , which in turn deterministically causes C , then B will be probabilistically independent of C conditional on A (since conditional on A , the values of all variables in the model are also determined). Additionally, faithfulness fails when there is failure of transitivity along a single path. In McDermott’s paradigmatic example (1995), a dog bites off a terrorist’s right thumb causing the terrorist to set off a bomb with his left hand instead (Fig. 2). The dog’s bite causes the terrorist to trigger the bomb in a different way, but makes no difference in the probability of an explosion. Accordingly, the dog bite is uncorrelated with the explosion even though the events are d-connected.

Discussions of whether it is probable that two paths will cancel out have no implications concerning failures of faithfulness not involving cancellation. Nevertheless, there is a principled reason for focusing on failures involving cancellation. Zhang and Spirtes (2008) make an important distinction between *detectable* and *undetectable* failures of faithfulness. A detectable failure of faithfulness is a case in which there is no model that is faithful to the probability distribution. McDermott’s case is one such example. It is straightforward to establish that no model satisfying the Causal Markov Condition is faithful to the distribution. Any model in which the middle node is a non-collider does not entail the unconditional independence of the dog bite and the explosion; the model in which the middle node is a collider does not entail the independence of the dog bite and the explosion conditional on a value of the middle node. The sense in which such failures are *detectable* is that one can use assumptions that are weaker than faithfulness to determine that no model is faithful to the distribution. In contrast, the failure involving cancellation depicted in

Fig. 1 Causal triangle with linear parameters

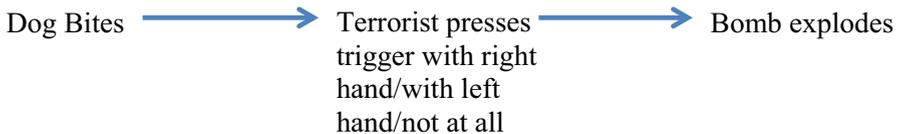
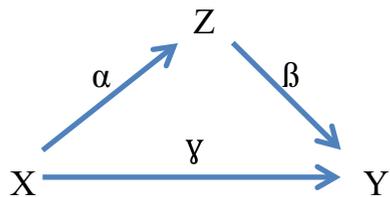


Fig. 2 Single-path failure of transitivity

Fig. 1 is an undetectable failure, since the model in which Z is a collider on a path between X and Y is faithful to the distribution. Unlike with detectable failures, where one can establish that there is no faithful model and thus avoid choosing false unfaithful models, with undetectable failures this strategy is not available.

Not all cancellations are undetectable failures of faithfulness. In fact, if one were to replace the model in Fig. 1 with a model including a mediator between X and Z along (what had been) the direct path or with an additional mediator between X and Y or between Y and Z , the failure of faithfulness resulting from cancellation would be detectable. Zhang and Spirtes introduce the concept of a triangle-failure of faithfulness, which is a failure involving three variables arranged into a triangle such that the two paths in the triangle cancel, and prove that if one assumes that there are no such failures, any other failure of faithfulness is detectable. This result provides some justification for the common practice of emphasizing triangle-failures to the exclusion of other types of failures.

In the following, I will only consider triangle-failures of faithfulness. Even if the Causal Faithfulness Condition turns out to be unjustified, the weaker assumption that there are no triangle-failures of faithfulness is independently important. Consider Pearl's (2009) minimality principle, which is weaker than faithfulness. This principle is entailed by the combination of CMC, triangle-faithfulness and a logically weak condition known as SGS-minimality (Zhang 2013). A clarification of the conditions under which triangle-faithfulness is justified thus has implications for this principle and for any principle that is logically stronger than it, such as frugality (Forster et al. forthcoming).

3 Debating the Probability of Cancellation

Spirtes et al. (2000, pp. 41–42) defend faithfulness on the grounds that the set of parameter values that cancel along the paths is infinitesimally small—measure zero—compared to the set of all assignments of parameter values. This result enables one to specify a set of probability distributions over possible parameter values for which the set of faithfulness-violating parameters values has probability zero. I will introduce their result with an analogy. Suppose I throw an infinitely thin dart at a dartboard and want to know the probability that I will hit a particular two-dimensional line. It follows from measure theory alone that the amount of space on the dartboard taken up by a two-dimensional line is infinitesimally small compared to the 3-dimensional surface of the board. More generally, an $n - 1$ dimensional area takes up Lebesgue measure zero of an n -dimensional space. This by itself does not entail anything about the *probability* that I will hit a particular line on the dartboard. To make claims about the probability of my hitting the line, we need to specify a probability distribution for my hitting different points on the board. Yet given that the dart and the line are infinitely thin, one only needs to make relatively weak assumptions about this probability distribution in order to get the result that there is zero probability of my hitting the line *exactly*.

Just as the space on a dartboard taken up by a point is infinitely smaller than the space on the board, the set of faithfulness-violating combinations of parameter

values is infinitely smaller than the set of all parameter values. SGS rely on the linear parameterization for the model in Fig. 1, for which cancellation occurs exactly when $\alpha\beta = -\gamma$. The set of parameter values that meet this constraint is measure zero compared to the set of all combinations of parameter values. Accordingly, for probability distributions on which parameterizations with measure zero are assigned probability zero, the probability of cancellation will be zero.

What can we say about the set of probability distributions for which the probability of cancellation will be zero? Steel (2006) shows that sets of parameter values of measure zero will have probability zero when “each parameter varies continuously considered separately and conditional on any subset of the others” (311). In other words, each parameter must vary in the distribution, the variation must not be limited to a discrete set of parameter values, and there must still be continuous variation in each parameter value when one conditions on others. This condition will be met when the parameter values are not perfectly coordinated. Given that such perfect coordination arguably never occurs,³ the measure zero argument can be used to show that the probability of two paths cancelling exactly is zero.

The measure zero argument is the most precise version of a family of arguments purporting to show that failures of faithfulness are improbable. Pearl (2009), for example, emphasizes that the parameter values that are needed to get cancellation are unstable in the sense that were they to have slightly different values, cancellation would not occur. In contrast, when faithfulness holds, the only variables that are conditionally independent are those that the model entails will be independent (for any set of parameter values). Glymour et al. (1987) similarly emphasize that models satisfying faithfulness entail the observed conditional independencies for *all* values of their free parameters. It is not clear whether these arguments in fact show that failures of faithfulness are *improbable* (later I argue that they do not). For the moment, I will focus on the measure zero argument, which explicitly provides the assumptions needed to conclude that they are.

As Steel (among others) notes, one shortcoming of the measure zero argument is that even if the set of *exactly* cancelling parameters is infinitely small compared to all sets, the set of *almost* cancelling parameters is not. Since one never has perfect data, near-failures of faithfulness are just as problematic as genuine ones. Moreover, Cartwright (1999) and Hoover (2001) argue that in certain scenarios, failures of faithfulness are in fact probable. In Steel’s example of such a scenario, a city plans to improve the roads, which would lead people to drive faster and consequently would increase the number of traffic fatalities. To avoid an increase in fatalities, the city hires more police officers to enforce the speed limit, but due to budgetary constraints they only hire enough police to cancel out the positive effect of

³ There is a slight ambiguity regarding the claim that perfect coordination of parameters is improbable. Is the claim that the parameters are coordinated a claim about the probability distribution over the parameters, or is it a claim about the process by which the parameters take on their values? If the former, then the claim that coordination is unlikely appears to be a restatement of the claim that faithfulness-violating parameter sets are rare, rather than an independent defense of the claim. If the latter, then we need to say more about the processes by which parameters take on their values. Here I do not dwell on this ambiguity, since these issues will be sorted out in what follows.

improved roads on traffic fatalities. In cases where there is an attempt to make two paths cancel out, we should expect failures of faithfulness. Let's refer to this response as the *policymaker objection*.

The policymaker objection, like the measure zero argument, makes a claim about whether certain combinations of parameter values are probable. Yet neither defenders nor critics of faithfulness explain what it means to assign a probability to a parameter. In general, causal models are paired with a probability distribution over the set of *variables* in the model. Given a model and this probability distribution, one can sometimes identify the parameters in the functional equations that relate a variable to its direct causes in the model. The parameters in the functional equations, however, do not vary. While it can be the case that two populations have different probability distributions over their variables and, correspondingly, have different parameter values, for any given probability distribution over the variables, the parameters will be constant. What, then, might a probability distribution over *parameters* be based upon? It clearly cannot be based on some sort of variation of the parameters in one's probability distribution, since parameters do not vary.

One might interpret the probabilities assigned to parameters as facts about an agent's degrees of beliefs. Yet, this is not how probabilities are typically interpreted in the causal modeling literature. Although there is little explicit discussion in the literature of how to interpret probabilities,⁴ in practice the probability distribution over the variables is estimated from data regarding the relative frequencies of property instances in a population and the parameters are identified given the combination of this distribution and a causal graph. Parameters indicate the functional relationships concerning how an effect variable will change given interventions on its causes, not merely an agent's beliefs about these relationships.

To my knowledge, Steel (2006) is the only writer to explicitly raise the question of how we should understand probability distributions over parameters. He claims that we should interpret such distributions as "physical chance process[es]" (308) and his discussion presupposes that policymakers are able to influence these chance processes. Nevertheless, he does not answer the question I have posed here regarding how to understand distributions over parameters, given that they do not vary. He does helpfully remark that although parameters are fixed within a population, they may vary across populations (309). But even if we grant that parameters can "vary" in the sense of differing across populations that share the same causal model, we need some account of how the physical chance processes responsible for variation in parameters across populations differ from those processes that are responsible for variation in variables *within* populations.

A quick point is in order about the term "parameters". In debates over faithfulness, it is common to talk in terms of linear parameters. Yet in using DAGs to make inferences from a probability distribution, one need not assume that the functional relationship linking an effect to its causes has any particular functional

⁴ Pearl (2009, p. 2) claims to interpret probabilities as subjective degrees of belief. Yet it is unclear how to reconcile this with his claim that the causal models that constrain these probabilities are objective features of the world.

form. For instance, the equation stating that a match lights (L) when it is struck (S) and there is oxygen in the room (O) is:

$$(\text{Match}) L = \max(S, O)$$

This equation does not include any parameters (i.e. coefficients). Yet were someone to make an assertion about the probability that these variables would have this relationship as opposed to some other, we could raise the same question about the basis for this probability assignment. Although I will talk about assignments of probabilities to parameters, the following discussion generalizes to other assignments of probabilities over different functional forms.

To summarize, both parties to the debate over cancellation make claims regarding whether cancellation is probable. To make sense of such claims, we need to clarify the basis for assigning a probability distribution to a set of parameters. In the following section, I provide such an account.

4 Where Parameters Come From

In Steel’s example illustrating the policymaker objection, hiring more policemen “shifts” the value of the parameter between road improvements and speeding. By considering what it means for a parameter to shift, we can learn what it is that determines the probability of a parameter in the first place.

Let’s begin with Steel’s example (Fig. 3a). R refers to road improvements, S to rates of speeding and T to traffic fatalities. There are two functional equations corresponding to this model, one that treats S as a function of R , and another that treats T as a function of R and S . This model does not necessarily include all causes of variables in the model. In Steel’s example, the effect of road improvements on speeding depends on the number of police that the city hires. The reason why one need not explicitly include these variables is that their influence is already reflected in the probability distribution over R , S and T . To see this, it helps to think of a case in which these variables are measured in multiple cities with different numbers of police and, consequently, differences in the effect of R on S (to simplify things, suppose that the only difference between the cities is the number of police hired). Although the effect of R on S will differ among the cities, the effect in the meta-

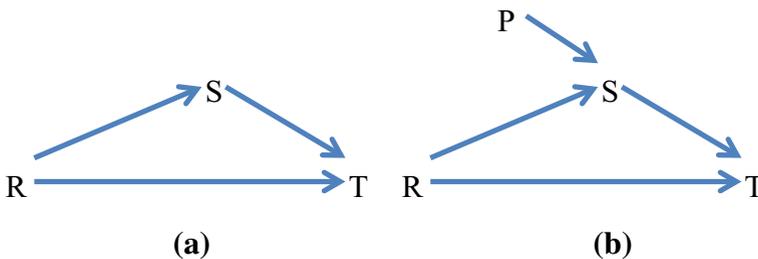


Fig. 3 a Graph for the policymaker scenario and b an augmented graph

population consisting of all the cities will be the average effect of R on S across the cities. To identify this average effect, one does not need to actually know the number of police hired in each city, since this is already reflected in the influence of R on S .

Of course, one *may* include a variable P for the number of police that the city hires (Fig. 3b). Then the functional equation for S will include P . The advantage of including P in one's model is that one can then determine the effect of R on S given different values of P . For short, we may refer to the probability distribution of S given an intervention on S and a particular value of P as the *P-specific* effect of R on S .⁵ Note that although the effect of R on S depends on the value of P , there is no variation in the functional relationships in either model. The model in Fig. 3a gives S as a fixed function of R . That in Fig. 3b gives S as a fixed function of R and P . Yet using this example, we may clarify what philosophers are talking about when they say that a parameter shifts. To talk of the parameter between R and S as shifting in the first model is to consider the way that the P -specific effect of R on S would vary in the second model. The sense in which the parameters in the leftward model “shift” is that a sociologist estimating the parameters of this DAG both before and after the hiring of police will get two non-equivalent estimates of the effect of R on S . Of course, the parameters do not actually shift—the sociologist is simply applying a model to two data sets. If she estimated the DAG for the *combined* data set, she would get a fixed estimate of the effect of R on S .

This example suggests a general strategy for discussing “variation” in parameter values. Variation in the magnitude of the effect of X on Y in one model may be understood as variation in the Z -specific effect of X on Y in an augmented model where Z is a cause of Y that interacts with X .⁶ In our example, talk of “variation” in the parameter relating R to S in (a) can be explicated in terms of the different P -specific effects of R on S in (b). I will refer to the strategy of interpreting the variation in parameters in one model as variation of variables in a broader model as the *augmented model strategy*.

Using this strategy, we can make sense of the common claim that the strengths of causal relationships can vary across populations. It is worth noting, however, that “populations” is not a technical term in the causal modeling literature. What the augmented model strategy enables one to do is to represent informal talk of populations formally. When populations differ in terms of some set of (possibly unmeasured) variables Z , populations differing in Z are characterized by the probability distribution over all non- Z variables that results from conditioning on that population's distribution for Z . Of course, we may not know the factors in

⁵ There are many possible measures for causal effects (Fitelson and Hitchcock 2011) and I do not here need to choose among them. For equations that treat an effect Y as a linear function of a cause X , it is natural to understand the effect of X on Y as the change in the expected value of Y given a one-unit change in X , and this is the sense in which I use the term in the main body of the text when referring to linear models. In areas where talk of causal effects may be eliminated in favor of a more general formulation, I provide said formulation in the footnotes.

⁶ More generally, for the probability distribution $P(Y|do(X), \mathcal{S})$ to “vary” between two populations P and P^* is for there to be some distinct variable Z with at least two distinct values z and z' such that $P(Y|do(X), \mathcal{S}, z) \neq P(Y|do(X), \mathcal{S}, z')$ and for populations P and P^* to differ with respect to the probabilities of z and z' . ($P(Y|do(X))$ refers to the probability distribution of Y given ideal interventions on X).

virtue of which populations differ. Yet to represent the difference between populations using causal models, we need to appeal to variation in some such factors.

Applying the augmented models strategy to the models that are presupposed by the policymaker objection reveals something interesting. To ensure cancellation, the city must not only hire police, but they must also consider the effects of improvements and speeding on traffic fatalities. One can model this using the graph in Fig. 4. In the graph, U_T is a variable representing the factors that make a difference in the effect of R and S on T . This variable influences not only T , but also the city's decision (D) regarding how many police to hire.⁷ U_T is therefore a common cause of the mediator and the outcome. It is significant that the scenarios envisaged by Cartwright and Hoover involve a common cause. When one can rule out such a common cause, there still might be cancellation within the RST triangle, but it would be *coincidental* in the same way that two uncorrelated variables might take on the same value.

The graph in Fig. 4 is not the only way that one might represent Steel's example. Perhaps the city does not begin with knowledge of the magnitude of the effects of R and S on T , but rather ensures that traffic fatalities do not increase using a method of trial and error. Specifically, they improve the roads, initially allow fatalities to increase, and then hire police officers until the fatality rate goes back to its previous level. We can represent this scenario using time-indexed variables (Fig. 5).

In Fig. 5, improving roads at time $t = 0$ (R_0) causes an increase in speeding at $t = 1$ (S_1) and these two variables have a net effect of changing the number of traffic fatalities at $t = 2$ (T_2). Supposing that the net effect is positive, the city will hire more police at $t = 3$ (P_3). This in turn will lower speeding at $t = 4$ (S_4) and reduce the rate of traffic fatalities at $t = 5$ (T_5). This process may be repeated until the rate of traffic fatalities at some later time $t = n$ equals its pre-policy level. Such tinkering will render R_0 and T_n probabilistically independent, producing a failure of faithfulness.

The scenario depicted in Fig. 5, while realistic, does not constitute a *triangle*-failure of faithfulness. Although there is a triangle connecting R_0 , S_1 and T_2 , the paths in this triangle do not cancel and R_0 and T_2 are probabilistically dependent conditional on any variable set. More generally, the failure of faithfulness in the example does not involve cancellation. Since the arguments we have been

⁷ There are several peculiarities about the proposed model that deserve mention. First, we may need a more complicated model to describe all of the causal relations that obtain. Most saliently, the city must know how hiring more police will influence speeding in order for them to ensure cancellation, so we may need to include a variable that is a cause of S (and D) into the augmented model. There may therefore be even more omitted common causes than that represented, but this would only strengthen my claim that the model omits common causes. Second, there are some complicated issues surrounding when we should include an agent's beliefs about her actions in a model. I take it is not *always* the case that if an agent intervenes on X in order to change Y and knows about the relationship between these two variables that we must represent how the agent learned this information. Yet the counterexample has two features that call for including this information in the form of an arrow from U_T to D . One is that the policymaker has knowledge of the parameter values. The other is that the parameters may have been different and that the policymaker is tracking these differences. As a result of these two features, the number of police hired does counterfactually depend on the functional relationships among the variables and should be represented.

Fig. 4 Augmented graph for policymaker scenario with parameter coordination

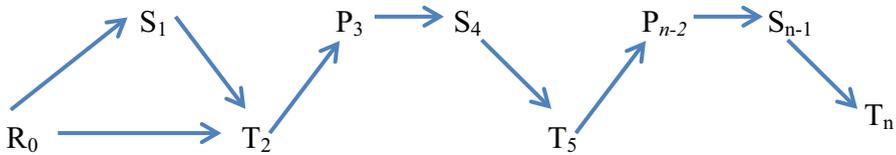
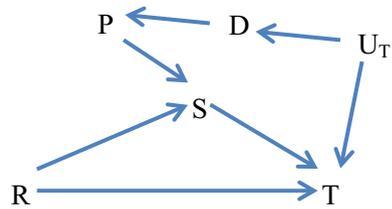


Fig. 5 Trial-and-error strategy for preventing increase in fatalities

considering concern the probability that the parameters in a model will cancel, the policymaker objection cannot rely on this scenario. It is for this reason that I represent the scenario using the graph in Fig. 4, in which the policymaker can genuinely be said to be coordinating the values of the parameters in the triangle.

Steel defends faithfulness against Cartwright and Hoover's counterexamples on the grounds that even when a policymaker seeks cancellation, cancellation is only ensured when two additional conditions are met: *selection* and *homogeneity*. Selection is the assumption that there is a process by which the policymaker can influence the parameters to make them tend towards the desired values. Homogeneity is the assumption that there are no factors that make the parameters randomly diverge from their desired values. Unless there is some degree of homogeneity, then even if the policymaker is able to get the parameter values to be close to their cancellation values, she will not be able to move them to precise values and her attempt to get cancellation will be unsuccessful.

We can use the augmented variable strategy to give Steel's selection assumption a causal interpretation. The scenarios involving selection of parameters may be represented in terms of the influence of common causes. Whether the homogeneity condition is met is determined not by the causal model alone, but rather by the probability distributions over the additional factors in the augmented model. Moreover, one must provide more information about these distributions than Steel does in order to use the presence or absence of homogeneity to draw conclusions about cancellation. Recall that Steel takes the measure zero argument to successfully rule out exact failures of faithfulness, and thus focuses his discussion on near failures. It is possible for the parameters to *nearly* cancel even if the parameters are allowed to vary somewhat from their cancellation values. Whether a failure of homogeneity leads to a failure of near-faithfulness depends both on precisely how much the parameters vary and on precisely how close to cancellation the paths must be in order to count as a near-failure of faithfulness.

Steel does not take selection and homogeneity to be necessary conditions for cancellation, but rather to be jointly sufficient. His main conclusion is a negative one: Cartwright and Hoover's examples do not yield near-failures of faithfulness without additional assumptions that might not be met. Accordingly, my claim that faithfulness can fail even in the absence of homogeneity does not undermine Steel's argument. The reason it is presently important to clarify Steel's selection and homogeneity assumptions is that both are informal ways of discussing features of the probability distribution over parameter values. As we have seen, the selection assumption can be characterized in terms of structural features of the augmented model—i.e. whether there is a common cause in the model—but whether the homogeneity assumption is satisfied depends on further features of the probability distribution. Below I suggest that the most promising avenues for defending faithfulness make assumptions only about the structural features of the augmented model, rather than on additional assumptions about the probability distribution over variables whose distributions are unconstrained by the structure. Were Steel to have provided an argument that in the absence of homogeneity failures of faithfulness are rare, this would be a way of defending faithfulness based on non-structural features of the augmented model. It is therefore important to clarify that Steel does not provide such an argument.

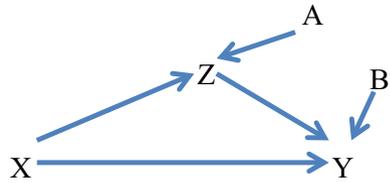
5 Non-coincidental Failures of Faithfulness

We have now seen that there is more going on in the policymaker example than meets the eye. While the counterexample is often represented using a simple triangle, a more careful description of the case using the augmented model strategy reveals that such a model omits a common cause. Moreover, causal sufficiency requires us to include such a variable, if there is one. Here I show that it is a general feature of non-coincidental failures of faithfulness that they require one to appeal to variables outside of the cancelling triangle. I then argue that when one does include these variables in the model, one only gets a genuine failure of faithfulness if there is cancellation for every value of these additional variables. Finally, I explain why it is implausible to think that this condition will be met in the standard counterexamples.

Consider the augmented model for Fig. 1 (Fig. 6). A is responsible for variation in the effect for the treatment on the mediator, and B is responsible for variation in the effects of the treatment and mediator on the outcome. If the values of A and B were probabilistically independent, then there would be no systematic relationship between the parameters that they influence. If A and B are dependent, then by the (contrapositive of the) CMC they must not be d-separated. Consequently, there are four ways that A and B could be related:

1. There is a chain of unidirectional arrows from A to B
2. There is a chain of unidirectional arrows from B to A
3. There is a common cause of A and B

Fig. 6 Augmented graph for Fig. 1



4. There is a collider on a path linking A and B and one has conditioned on the collider (or on one of its descendants)

If any of the first three possibilities obtain, then the mediator and the outcome will share a common cause. The fourth possibility is worth considering further.

Andersen (2013) notes that faithfulness can fail in biological systems in which certain quantities are maintained at an equilibrium value. Consider the effect of running on body temperature. Running increases body temperature, but also causes one to sweat, which reduces body temperature. Sweating evolved as a means of maintaining body temperature within a constant range, so the two causal paths between running and body temperature will roughly cancel out. A simplified story of how this evolved is that organisms in which these paths did not cancel did not survive. Intriguingly, within a DAG the evolution of balancing mechanisms is modeled as a case of selection bias.

The parameters in a non-augmented model for running, sweating, and body temperature, can be represented as causes in an augmented model (the sweating and temperature mechanisms in Fig. 7). These two causes jointly influence survival, since the relationship between these parameters makes a difference for whether the organism survives. The survival variable is therefore a collider on a path between sweating and body temperature. Think of *survival* variable as a dichotomous variable taking on the values yes or no. When scientists look only at living organisms, they effectively condition on this variable at the value of $survival = yes$, producing selection bias. There is a “correlation” between the magnitudes of the effects along each path within the triangle because only organisms in which these magnitudes balanced made it into the sample.

While I have evaluated only two scenarios, the argument for the claim that non-coincidental cancellations require there to be variables outside the triangle is

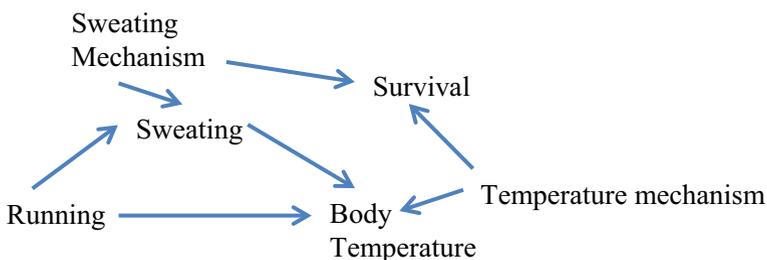


Fig. 7 Graph for a homeostatic system

straightforward. First, we assume that cancellation involves coordination of parameters. A triangle corresponds to two functional relationships and unless these relationships were somehow coordinated, cancellation would be coincidental. Second, the way to understand this coordination is in terms of a correlation between the variables in the augmented model. Third, by CMC there must be some causal relationship between correlated variables. When one spells out the possible relationships, each involves either a common cause of variables in the triangle or a common effect that one conditions on. Causal sufficiency requires us to include the common cause, and we would need to include the common effect in order to avoid selection bias. So common modeling assumptions require us to include variables from outside the triangle.

What happens when we include these additional variables? In cases where cancellation is due to selection bias, it is straightforward to see that when one includes the common effect and does not condition on it, there is no longer a failure of faithfulness. The manner in which selection bias explains the cancellation is precisely that one only observes the individuals for whom the paths cancelled. So in the population consisting of organisms who survived and those who did not there will not on average be cancellation. In cases where the cancellation is due to a common cause, things get a bit more complicated, since we need to say more about the possible values of the common cause variable.

Consider again Fig. 4, which includes the variable D corresponding to the decision that the policymakers make regarding how many police to hire. If one possible decision that the policymakers could make is to do nothing, then clearly for that value of D any cancellation would be coincidental. In the absence of such a coincidental cancellation, road improvements will be unconditionally associated with traffic fatalities. Accordingly, for the variables in the augmented model, road improvements would be unconditionally associated with traffic fatalities and there would not be a triangle-failure of faithfulness.⁸ Similarly, if there are some values for D such that the number of policemen the city hires does not lead to cancellation, there will not be a failure of faithfulness. So in order for there to be a failure of triangle-faithfulness in the augmented model, it must be that there is cancellation for every value of D .

What would it mean for D 's values to be exhausted by those that ensure cancellation? Causal variables, qua variables, represent a jointly exhaustive partition over the space of possible variable values. So to say that all of D 's values ensure cancellation is to say that it is in some sense not possible for D to take on a value that does not ensure cancellation. The sense of possibility relevant to a certain variable is usually left implicit. A variable corresponding to whether patients in a study took 0, 1, or 2 pills presumably refers to the different possible treatments one could receive in the study, and does not indicate that someone outside the study

⁸ One might suppose that there would still be a triangle-failure of faithfulness, since the true model would not entail that road improvements and fatalities are independent conditional on the relevant value of D . Yet in order to satisfy faithfulness, a model only needs to entail full conditional independencies, not partial independencies. That is, failures of faithfulness are cases in which X is independent of Y conditional on Z for *all* values of Z (rather than just some) and the model does not entail this set of conditional independencies according to CMC.

context could not take three pills. Without providing a general principle for how to specify the values of a variable, we might say (vaguely) that variables whose values do not exhaust the logical space should have variables that cover every live possibility. This idea is in the same spirit as the positivity assumption, which requires one to have data corresponding to every combination of variable values. Positivity is needed to learn about the relationships among variables under varying counterfactual circumstances. While positivity does not say anything explicit about how to choose one's variables, it characterizes one sense in which DAGs require one to have data from a sufficiently rich range of counterfactual circumstances.

These considerations motivate the suggestion that the variable D should only have values that ensure cancellation if other values are not live possibilities. If so, then in order to get failures of faithfulness in the augmented model, it needs to not merely be the case that there is cancellation, but that the cancellation be in some sense necessary. It is unclear if there are such cancellations, but even if there are, it is a reasonable rule of inference to suppose that there are none in the absence of evidence to the contrary. Now, it may initially seem that the reason we think that necessary cancellations are implausible is the familiar idea that were some agent to aim for cancellation, she would probably not succeed. But it is in fact an independent reason. Even if an agent were extremely reliable at ensuring cancellation, this would not render the cancellation necessary, provided that there was a live possibility of the agent not trying to make the paths cancel. What makes it plausible that there are not necessary cancellations is not that one has a hypothesis about the causes of cancellations such that one can assign a probability to the claim that the paths will in fact cancel. It is the idea that when paths cancel (for whatever reason) they might not have done so.

The assumption of no necessary cancellation has implications for both coincidental and non-coincidental failures of faithfulness. This assumption ensures that whenever there is a population for which there is cancellation, there is a broader (possible) meta-population representable by an augmented graph in which faithfulness does not fail (since the meta-population includes populations in which the paths do not cancel). It is important to see, however, that my defense of faithfulness in the present section is not merely that when faithfulness fails in one population, it does not fail in some meta-population. To say so would be to change the subject, since faithfulness *does* fail in the initial population. Rather, my argument relies on there being cases in which we are required by standard modeling assumptions to use the DAG corresponding to the meta-population. Namely, with non-coincidental failures we need to use the augmented model in order to satisfy causal sufficiency and avoid selection bias. In short, I argue not merely that there exist augmented models in which faithfulness obtains, but that these are models that standard assumptions require us to use.⁹

Independent of the question of whether we should assume away necessary cancellations in general, it is clear that the stated counterexamples to faithfulness do not involve such cancellations. Both the policymaker and homeostatic systems counterexamples explain particular cancellations in terms of processes in whose

⁹ I am grateful to an anonymous reviewer for pushing me on this point.

absence cancellation would not be ensured. A policymakers' decision to try to make causal paths cancel is based on the perception that in the absence of such a policy, the paths would not cancel. The selection explanation assumes that in the population for which there is no selection bias there would not be cancellation. Since these examples by their nature do not involve necessary cancellations, there will not be a failure of faithfulness in the distribution that includes the variables from outside the triangle. And we must include these variables to satisfy causal sufficiency and avoid selection bias.

Of course, in practice we often have causally insufficient variable sets and cannot always avoid selection bias. If certain counterexamples to faithfulness involve violations of causal sufficiency and no selection bias, this might be just taken to show that violations of these other principles are more widespread than we thought. The importance of the link between non-coincidental failures of triangle faithfulness and these other principles is not that we can use them to rule out such failures a priori. Rather, the link renders our reasoning about whether faithfulness holds in a given domain to be continuous with our more general qualitative reasoning about whether we have enough variables to avoid confounding. On the face of it, the principles used to avoid confounding play no role in avoiding failures of faithfulness, since there may be failures even in the absence of confounding. Yet because avoiding confounding requires us to include additional variables that expand the range of possibilities that we consider, the conditions that prevent confounding play a role in ruling out certain types of failures of faithfulness. What might appear to be a failure of faithfulness relative to a limited variable set will disappear when one considers the wider range of possibilities.

The assumptions of causal sufficiency, no confounding, and no necessary cancellation are only able to rule out *non*-coincidental triangle-failures of faithfulness. They leave open the possibility that there might be cancellation by coincidence. We now turn to the question of what one should say about such coincidental failures.

6 Coincidental Failures of Faithfulness

The claim that a failure of faithfulness is coincidental does not entail that it is improbable. To see why, consider two coins with a bias of .9 towards heads. It is probable that both coins will land heads, though this outcome would be coincidental in the sense that I've been using the term. Since saying an outcome is coincidental does not entail that it is improbable, the assumptions that rule out non-coincidental failures do not license the conclusion that failures of faithfulness are improbable. This may seem disappointing, given that defenders of faithfulness have aimed to show that failures of faithfulness are improbable. In this section, however, I argue that defenders of faithfulness have drawn too strong a link between the question of whether faithfulness is justified and that of whether it is likely to fail. There are explanatory and pragmatic bases for preferring models that do not involve coincidental failures.

How we should think about coincidental failures depends in part on how much knowledge we have regarding the system in question. At one extreme, we have sufficient knowledge regarding the probability distributions over the variables in the augmented model to know that the paths will cancel. To have such knowledge, one would presumably already have to know what the true model is and there would be no need for faithfulness. At the other extreme, we have no knowledge whatsoever of the probability distribution over the parameters. Between these extremes, there are domains in which one might have some beliefs about the stability of the causal relationships. For example, if one is seeking to determine which causal relationships obtain among various psychological and behavioral variables, one may reasonably assume that whatever the relationships are they will vary across contexts and individuals. The reasoning here is that the strengths of these relationships depend on a wide range of independent factors that vary from person to person. In this type of case, the natural thing to say is that although it is certainly possible for one to get the wrong model as a result of a failure of faithfulness, by looking at the same set of variables over a variety of contexts, one could discover one's mistake.

The idea that we can detect failures of faithfulness when parameters vary over time or across populations is by no means new, though its implications for faithfulness have remained obscure. Pearl, for example, plausibly suggests that algorithms whose reliability is guaranteed given CFC might be more reliable when applied to data from longitudinal studies (2009, p. 63). Yet, right before saying this, he uses parameter variation to argue that failures of faithfulness will be rare. He relies on Aldrich's (1989) notion of *autonomy*—the idea that particular structural equations in a causal model are invariant to changes in other structural equations. He writes that autonomy entails that the parameters in the equations:

[C]an and will vary independently when experimental conditions change. Consequently, equality constraints of the form $\alpha = -\beta\gamma$ are contrary to the idea of autonomy and will rarely occur under natural conditions. (63)

Using the augmented model strategy, we can clarify the notion of autonomy. The parameters in different equations can vary independently in the sense that the variables corresponding to the parameters in the augmented model will be uncorrelated. Yet, contra Pearl, this does not guarantee that failures of faithfulness will be rare, only that they will be coincidental.

Pearl conflates distinct ways of defending faithfulness: the conclusion that failures are rare is distinct from the conclusion that failures of faithfulness will be detected in the long run. The conclusion that failures of faithfulness are extremely rare suggests that there's no need to worry about them. Yet the conclusion that faithfulness will be detected in the long run is compatible with there being failures of faithfulness that mislead us in the short run. The fact that if faithfulness fails in one study we may discover our error in later ones does not by itself have any implications for whether we should take faithfulness to be a reliable assumption in any particular study. It is nevertheless an important property of faithfulness that under the relevant conditions it will not permanently mislead us. Practically, this feature of faithfulness provides an additional reason to model a variable set over distinct populations and varying contexts.

What *should* we say about whether we are justified in assuming faithfulness in the short run? This question arises both in the scenario where one makes a general assumption that the parameters vary and in one in which one truly has no beliefs about the probability distribution over the parameters. In such cases, one cannot show that failures of faithfulness are rare, but faithfulness may nevertheless be provided with a defeasible justification on explanatory and predictive grounds, as follows.

First, one can defend faithfulness on the grounds that faithful models provide better explanations. When there is a coincidental failure of faithfulness, there is no explanation for why the parameters in the model have the values they need to have in order to cancel. Faithfulness may thus be defended on the grounds that we should prefer models that do not posit such brute facts. This suggestion is similar to the Lange's (1995) defense of a predecessor of faithfulness. The problem with Lange's account is that he is unable to provide compelling grounds for distinguishing between the parts of a model or theory that do and do not require explanation. The present discussion provides grounds for saying that coordinated parameters require explanation. Relationships between parameters are ordinarily explained either by a common cause or selection bias. While we cannot rule out the possibility that parameters will cancel by coincidence, we can nevertheless penalize models that treat as a brute fact something that ordinarily can be explained.

An alternative defense of triangle-faithfulness rests on the distinction between prediction and accommodation. According to this defense, while faithful models predict all conditional independencies in the distribution, unfaithful models merely accommodate them, since the conditional independencies only follow from the model when it is combined with assumptions about the values of the parameters. Faithfulness can thus be defended on the grounds that faithful models make more predictions than unfaithful ones and should thus be rewarded when these additional predictions are validated (cf. Popper 1959).

This second defense of faithfulness assumes that we should compare models with their parameters unspecified. If one is allowed to specify the values of the parameters in the unfaithful model, then the parameterized version of the unfaithful model entails the same conditional independencies as any parameterized version of the faithful model. Cartwright (1989) has criticized faithfulness (and causal models more generally) on the grounds that there is no basis for comparing models with their free-parameters unspecified. Why think that the structure of a model is in some sense more fundamental than the quantitative relationships among the variables? In Cartwright's words:

It makes sense to look exclusively at causal structures (i.e., their graphs) only if one assumes that...any theory that implies the data from the causal structure alone is more likely to be true than one that uses the numbers as well. (1989, 76–77)

The distinction between the qualitative and quantitative aspects of a model is not essentially connected to faithfulness, but is already present with CMC. While the DAG alone entails the conditional independence relations, relationships between parameters in the model need to be written in by hand. This does not in any direct

sense show that models that entail the conditional independencies from their structures alone are more likely to be true, but it does provide a basis for rewarding models that are able to predict conditional independencies without the aid of additional, independent, assumptions about the values of parameters.

In scenarios where one has reason to believe that failures of faithfulness are rare, the proposed defenses are unnecessary. If, in contrast, failures are known to be common, the explanatory or predictive virtues of faithful models will be of little consolation. The proposed defenses are useful in precisely those cases where we can rule out non-coincidental failures of faithfulness and do not have adequate knowledge to make a precise statement about the probability of cancellation. Yet even in the absence of such probabilistic information, we may defend the application of faithfulness to these cases on explanatory and predictive grounds.

7 Conclusion

A peculiarity regarding debates over faithfulness is that although it is widely understood to be an epistemic principle for choosing between models based on a probability distribution, debates about its validity have emphasized the metaphysical question of how often faithfulness is true. By using the augmented model strategy, we can shift the terms of the debate from questions about how often faithfulness fails to questions about its basis as a principle for selecting models. What emerges from our discussion is that many of the purported counterexamples need not be ruled out using precise knowledge about the probability of cancellation, but are already ruled out by standard causal modeling assumptions combined with the implicit assumption that if there is a cancellation, it is not a necessary one. The reason this is significant is not that the assumptions of causal sufficiency and no selection bias are always known to obtain. Rather, it reveals that a defense of faithfulness need not rely on stipulations about the probability of failure, but rather on the same type of qualitative causal knowledge we use more generally in causal inference. In a similar vein, in the absence of probabilistic knowledge one may use explanatory or predictive criteria justify the assumption there will not be coincidental cancellations. By distinguishing between the different forms of cancellation, I have both clarified prior debates over faithfulness and suggested further ways of defending it.

Acknowledgements I would like to thank Dan Hausman, Jonathan Livengood, Erik Nyberg and Jan Sprenger for reading prior drafts and providing helpful comments. This paper benefited from questions I received when presenting versions of it to audiences in Wisconsin, Kansas State and at the 2014 meeting of the Eastern APA in Philadelphia. It especially benefitted from the comments of Bruce Glymour, Reuben Stern, Elliott Wagner, Elliott Sober, Hayley Clatterback and David O'Brien. Finally, I am grateful to two anonymous reviewers for their insightful feedback.

Funding This work was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, 41, 15–34.
- Andersen, H. (2013). When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, 80(5), 672–683.
- Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford: Clarendon Press.
- Cartwright, N. (1999). *The dappled word: a study of the boundaries of science*. Cambridge: Cambridge University Press.
- Elwert, F., & Winship, C. (2014). Endogenous selection bias: the problem of conditioning on a collider variable. *Annual Review of Sociology*, 40, 31–53.
- Fitelson, B., & Hitchcock, C. (2011). Probabilistic measures of causal strength. In P. M. K. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 600–627). Oxford: Oxford University Press.
- Forster, M., Raskutti, G., Stern, R., & Weinberger, N. (forthcoming). The frugal inference of causal relations. *British Journal for the Philosophy of Science*.
- Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structures*. New York: Academic Press.
- Hoover, K. D. (2001). *Causality in macroeconomics*. Cambridge: Cambridge University Press.
- Lange, M. (1995). Spearman's principle. *British Journal for the Philosophy of Science*, 46, 503–521.
- McDermott, M. (1995). Redundant causation. *British Journal for the Philosophy of Science*, 40, 523–544.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.
- Popper, K. (1959). *The logic of scientific discovery*. London: Routledge.
- Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search* (2nd ed.). New York: Springer.
- Steel, D. (2006). Homogeneity, selection, and the faithfulness condition. *Minds and Machines*, 16, 303–317.
- Zhang, J. (2013). A comparison of three Occam's Razors for Markovian causal models. *The British Journal for the Philosophy of Science*, 64(2), 423–448.
- Zhang, J., & Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 7, 239–271.