Synthese (2025) 206:204
https://doi.org/10.1007/s11229-025-05268-0

ORIGINAL RESEARCH

Causal inference and the metaphysics of causation

David Papineau[1]

Received: 4 February 2025 / Accepted: 10 September 2025

Abstract
The techniques of causal inference are widely used throughout the non-experimental sciences to derive causal conclusions from probabilistic premises. This poses a philosophical question. What in the nature of causation accounts for the possibility of causal inference? In this paper I shall seek to answer this question. I shall develop a metaphysical theory of causation designed to explain the success of causal inference techniques. I shall then indicate how this theory also accounts naturally for single-case causation and the relation between causation and rational action. Finally, I shall compare the resulting picture with "interventionist" approaches to causation.

Keywords Causation · Causal inference · Causal Bayesian networks · Interventionism · Structural equations

1 Introduction

The techniques of causal inference are widely used throughout the non-experimental sciences to derive causal conclusions from probabilistic premises.

This poses a philosophical question. What in the nature of causation accounts for the possibility of causal inference?

In this paper I shall seek to answer this question. I shall develop a metaphysical theory of causation designed to explain the success of causal inference techniques. I shall then indicate how this theory also accounts naturally for other aspect of causation.

In more detail, Sects. 2–5 below will outline the techniques of causal inference. Sections 6–10 will show that "probabilistic theories of causation" cannot satisfac-

✉ David Papineau
david.papineau@kcl.ac.uk

[1] Department of Philosophy, King's College London, London WC2R 2LS, UK

Published online: 07 October 2025

torily account for their success. In Sects. 11–13 I shall develop my own positive explanation in terms of *recursive structures of laws with exogenous independence*—"RLIs" as I shall call them. Section 14 will then be about single-case causation, Sect. 15 about the connection between causation and rational action, and Sect. 16 about indeterminism. Finally, Sects. 17–19 will offer some thoughts about "interventionist" accounts of causation.

2 Causal inference

Let us assume that we know the joint probability distribution over some set of macroscopic quantities {**V**}—in the sense of an objective probability distribution that obtains within some background field defined by fixed values for certain other quantities.

For example, we might know the joint probability distribution of weight, hypertension and anxiety levels in a background field defined by 21st century British residence, adulthood, and no long-term comorbidity.

Now, the techniques of causal inference often allow us to draw causal conclusions from such probability distributions. For example, suppose that the above probability distribution tells us that anxiety and weight are uncorrelated but that both are correlated[1] with blood pressure. Then the causal inference techniques tell us (assuming "causal sufficiency"—I'll come to that in a moment) that anxiety **A** and weight **W** must be causally independent of each other, but that both exert a causal influence on hypertension **H**. (At intuitive first pass, the correlations between **H** and **A** and **W** respectively mean **H** must be causally connected with both **A** and **W**; but if it were connected to them other than via the structure below, we would expect **A** to be correlated with **W**, which it isn't.)

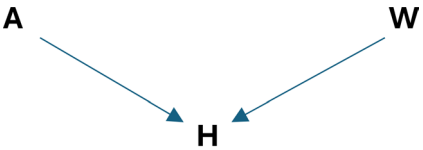


Diagram 1

3 Bridge principles

Can we state the principles behind inferences like this? By their nature, such principles must bridge the gap between correlational facts and causal facts. So I shall call them "bridge principles".

[1] When I speak of quantities being *correlated*, I simply mean that their *probability distributions are not independent*. In statistical circles, correlation is sometimes understood more strictly, as a measure specifically of *linear* relatedness. Still, this strict reading is not required by everyday usage, and my exposition will proceed more smoothly without it. (So in my terms two quantities can be correlated even if their Pearson correlation coefficient is zero.)

We can usefully view causal inference as hinging on two such bridge principles—one that takes us from correlations to causal connections, and another that takes us from the absence of correlations to the absence of causal connections.

In the literature on causal inference, these two principles are normally formulated as the ***Causal Markov Condition*** and the ***Faithfulness Condition*** (Spirtes et al., 1993; Pearl, 2000). However, I find it more illuminating to proceed slightly differently. I shall stick with the Faithfulness Condition, but will replace the Causal Markov Condition with a claim I shall call ***Reichenbach's Principle***.

Before articulating these principles it will be helpful to introduce a couple of definitions.

First, I shall say that two quantities **X** and **Y** are ***causally linked*** if **X** causes **Y** (possibly indirectly via intermediaries), or **Y** causes **X** (again possibly indirectly), or **X** and **Y** have a (possibly indirect) common cause—but not if **X** and **Y** only have a common effect.

Second, I shall say that some quantity **W** ***screens off*** the correlation between **X** and **Y** if $\text{Prob}(\mathbf{X}\&\mathbf{Y}) \neq \text{Prob}(\mathbf{X})\text{Prob}(\mathbf{Y})$ but $\text{Prob}(\mathbf{X}\&\mathbf{Y}|\mathbf{W}) = \text{Prob}(\mathbf{X}|\mathbf{W})\text{Prob}(\mathbf{Y}|\mathbf{W})$. (In such cases I shall also say the correlation between **X** and **Y** disappears when we "control" for **W**. Note that "controlling" in this sense does not require experimental manipulation, active intervention, or anything like that. It's simply a matter of attending to the relevant conditional joint probability distributions.)

Now for ***Reichenbach's Principle***. This says:

> If two quantities **X** and **Y** are correlated, then they must be causally linked.
>
> More generally[2], if two correlated quantities **X** and **Y** are not screened off by the quantities in some set {**Z**}, then **X** and **Y** must be causally linked by intermediaries that do not involve {**Z**}.[3]

As will be apparent, Reichenbach's[4] Principle takes us from correlational claims to causal claims. The Faithfulness Condition takes us in the opposite direction, from

[2] As will be seen, the first part of Reichenbach's Principle as stated is a special case of the second part. (If an **X**-**Y** correlation isn't screened off by the empty set, then **X** and **Y** must be causally linked full stop.) Still, it is easier to follow things if the special case is presented first.

[3] As stated, Reichenbach's Principle is open to counter-examples that arise because controlling for a "collider"—a quantity with two causes, like **H** in diagram 1—will induce a negative correlation between the causes. To deal with this we need the notion of "d-separation" (Pearl, 1988). Once we have this we can restate Reichenbach's Principle as follows: if two correlated quantities **X** and **Y** are not screened off by the quantities in some set {**Z**}, then **X** and **Y** cannot be *d-separated* by {**Z**}. I shall by-pass these complications in the interests of a smoother exposition.

[4] The name is inspired by Hans Reichenbach's pioneering work in *The Direction of Time* 1956. It is unclear whether Reichenbach himself actually endorsed this Principle. He was certainly committed to the "Principle of the Common Cause"–if two quantities are correlated, then one must be causing the other or they must have one or more common causes, and in the latter case controlling for the common causes will screen off the correlation. In addition, Reichenbach was aware that, in typical cases where **X** causes **Y** through {**Z**}, the **X**-**Y** correlation will be screened off by {**Z**}. But he did not, as far as I know, try to put all this together in a general claim.

causal claims to correlational claims—or, equivalently, from the negation of correlational claims to the negation of causal claims.

The ***Faithfulness Condition***:

> If two quantities **X** and **Y** are not correlated, then they are not causally linked.
>
> More generally[5], if the correlation between **X** and **Y** is screened off by the quantities in some set {**Z**}, then **X** and **Y** are not causally linked by intermediaries that do not contain any of {**Z**}.

4 Causal sufficiency

Now, as we shall see, there is plenty of room to query whether our two bridge principles are unexceptionally true. Still, unless we are willing to dismiss as groundless a vast body of well-respected research across a range of disciplines, we need to accept that they contain at least some element of truth. So for the moment I shall take them as given, and explore their ability to deliver causal conclusions. We shall address the exceptions and qualifications in due course.

As it happens, we need to add a further assumption to Reichenbach's Principle and the Faithfulness Condition before we can draw any definite causal conclusions. We need also to assume that the set of quantities {**V**} we are dealing with is ***causally sufficient***, in the sense that it contains any proximal common causes of quantities that it does contain.

To see why, return to our initial example of anxiety **A**, weight **W** and hypertension **H**. When I inferred their causal relations from their correlations, I slipped past a potential problem. Suppose we now unpack the reasoning used to reach Diagram 1 as follows. Focus on the **A**-**H** relation and apply Reichenbach's Principle. Since **A** and **H** are correlated, they must be causally linked; and, since their correlation isn't screened off[6] by anything in {**V**}, this link must be due either to **A** causing **H** directly or vice versa. And then, since **A** and **W** are probabilistic independent, the Faithfulness Condition tells they can't be causally linked, which leaves Diagram 1 as the only possible causal structure.

The hole in this reasoning is that we only know that the **A**-**H** correlation isn't screened off *by anything in {V}*, not that it isn't screened off by any quantities at all. And this leaves us hostage to the possibility that some common cause *outside the study* might yet screen off the **A**-**H** correlation, thus undermining the claims that one of **A** or **H** must be causing the other.

To illustrate, imagine, not implausibly, that social class **C**, is a common cause of both anxious feelings **A** and hypertension **H** but that neither of the latter causes the

[5] Again, the first part is a special case of the second part.

[6] For **W** to screen off the **A**-**H** correlation it would have to be correlated with **A**, which by hypothesis it isn't.

other. Then the correlations within {**V**} would have been exactly the same, yet we would have been wrong to infer that **A** causes **H**. Rather the causal structure would have been:

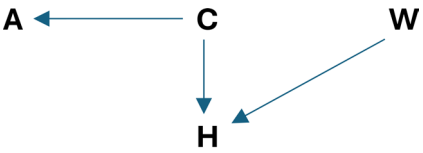


Diagram 2

It is worth noting that this problem relates specifically to the omission of *common* causes from a study, not to the omission of *any* causes of the quantities in question. Suppose that some omitted **Z** does in fact screen off the correlation between two quantities **X** and **Y** included in some {**V**}, not because it is a common cause of them, but because it is causally intermediate between them, with **X** causing **Y** via **Z**, or **Y** causing **X** via **Z**. In that case, any conclusion that **X** caused **Y**, or vice versa, arrived at on the basis that the **X**-**Y** correlation is not screened off by anything in {**V**}, would not thereby be invalidated, in the way it would be if **Z** was a common cause of **X** and **Y**.

True, in such a case a causal link that looked direct in relation to {**V**} alone would in truth only be indirect, proceeding via the intermediate cause **Z**. But of course that is just what we would have expected anyway. After all, it is natural to suppose that causation is *dense*, in the sense that in reality, as opposed to in our selective study, there will always be causal intermediaries whenever some **X** causes a distinct **Y**, and so that in reality there will never be any absolutely direct causal links.

This point is crucial to the practical business of causal inference. Empirical researchers need to attend to common causes, but not to all causes. Since any quantity will typically be subject to an indefinite number of causal influences, a requirement to attend to *all* causes of the quantities under study would render practical causal inferences impossible. By contrast, the requirement that we include all *common* causes is far less demanding and by no means obviously unsatisfiable.

This is not to say that this requirement is trivial. The possibility of hidden "confounders" is always a danger for causal inference. Different disciplines tend to have different attitudes to this danger. Roughly speaking, epidemiologists and social scientists hold that it can often be surmounted by a judicious selection of surveyed quantities. Enthusiasts for "evidence-based medicine", by contrast, feel that observational surveys can never eliminate the danger of hidden confounders, and that only "randomized controlled trials" can justify causal conclusions. Econometricians tend to lie somewhere in between.

These disciplinary differences are largely orthogonal to my central concerns. My focus is on the metaphysics of causation, not on the methodology of causal research. So I do not need to take a view on how best to overcome the practical danger of confounding common causes. For me the important point is simply that we can use the bridge principles to infer definite causal conclusions once we assume that the quantity sets that they are applied to are causally sufficient, in the sense of including any common causes of quantities they do include. This hypothetical claim is granted

even by those who are sceptical about eliminating the threat of hidden confounders, and it alone poses the challenge of explaining what in the nature of causation warrants such causal inferences.

5 Determining causal structure

Of course, not any set of correlations over any {**V**} will allow the bridge principles plus causal sufficiency to uniquely determine the causal relations between the quantities in {**V**}.

For example, if we are just given that **X** and **Y** are correlated, then this will leave us undecided whether **X** is causing **Y** and vice versa.

Again, if we are given that **X** and **Y** are correlated and that their correlation is screened off by **Z**, then these three possibilities remain open:

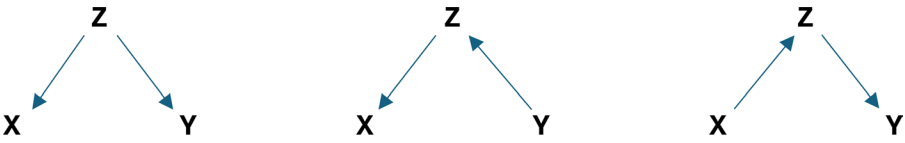


Diagram 3

Still, it is provable (Spirtes et al., 1993 Theorem 4.6) that in all such cases, where the correlations between the quantities in some set **{V}** are insufficient for the bridge principles plus causal sufficiency to determine a unique causal order, there will be further possible correlations, perhaps involving further possible quantities, that will so determine a unique causal order. (To see why, note that whenever it is indeterminate whether **X** causes **Y** or vice versa, some **Z** that is correlated with **Y** but not **X** would show **X** causes **Y**, while some **Z** that is correlated with **X** but not **Y** would show the converse. In this connection, it might be helpful to look at Diagram 1 again.) Of course, it is a further question whether this possibility will always be actual—that is, whether reality will always provide further quantities and correlations so to resolve causal indeterminacies. I shall return to this issue below.

6 Probabilistitic theories

So the bridge principles are capable of determining a causal order given a rich enough pattern of correlations over a causally sufficient set of quantities.

This now brings us to a fundamental question. What grounds the possibility of causal inference? Equivalently, why are the bridge principles true (to the extent that they are)?

One possible answer is that this is simply a contingent matter. Causal structures are one thing, their correlational signatures another. Systematic observation has shown us that causal structures are typically associated with certain correlational patterns, but it didn't have to be like that. That's just how it pans out in in the actual world.

(Thus Judea Pearl is wont to say, perhaps with tongue in cheek, that the possibility of causal inference is "a gift from the gods"—2018 116.)

Well, that is one possibility, but it would be explanatorily more satisfying if we could somehow show how the bridge principles are consequences of the nature of causation itself. I shall now consider two attempts to show this. In this section I shall discuss "probabilistic theories of causation". These seek to identify causal structures directly with the correlational patterns which display them. I shall demonstrate that this does not work. Then in the next section I shall show that we can do better by reducing causation instead to systems of directed "structural equations".

The first probabilistic theories of causation were proposed in the middle of the twentieth century by I.J. Good (1961-2) and Suppes (1970). They were various difference between their proposals, but at bottom they were both variations on this idea:

> A temporally earlier quantity **X** causes a later quantity **Y** if and only if they are positively correlated and this correlation is not screened off by any yet earlier set of quantities {**Z**}.

Note how this formulation appeals to temporal order in analysing causal order. However, more recent analysis of the logic of causal inference has made it clear that this is unnecessary—the bridge principles need no assumptions about temporal order to infer causal structure from sufficiently rich sets of correlations.

On this basis, a number of more recent writers have urged versions of a probabilistic theory of causation that avoid any appeal to temporal direction. Causal order is nothing but the structure implicit in sufficiently rich sets of correlations. Versions of this view have been endorsed by Spohn (2001), Glymour (2004), and Papineau (1992, 2001).

Is not this reductionist project undermined by the fact that the bridge principles can only determine causal structure given *causally* sufficient sets of quantities? But this threat of circularity can be finessed. Whenever causal insufficiency leads the bridge principles to generate mistaken causal conclusions, these mistakes will be exposed by adding the omitted common causes to the quantity set. This means that the probabilistic project can reductively equate the true causal order with those patterns that, so to speak, emerge in the limit—that is, with those causal patterns that do *not* get overturned when we expand initial quantity sets by bringing in further quantities.

7 An explicit probabilistic definition

At the end of the Sect. 5 I noted there is no guarantee that reality will always provide enough correlations to resolve every causal indeterminacy. Interestingly, though, if we do assume that it will, then it becomes possible to frame an explicit probabilistic reduction of causation[7]:

[7] This definition is inspired by Daniel Hausman's analysis of causation in his *Causal Asymmetries* 1998; it should be noted that he himself does not frame his analysis in terms of probabilities because of concerns

(1) **A** causes **B**.

if and only if.

(2) (a) **A** is correlated with **B**, and (b) everything correlated with **A** is correlated with **B**, and (c) something correlated with **B** is not correlated with **A**.

It is not hard to prove that this equation follows from the bridge principles, once we assume that every quantity is associated with some independent source of variation—more precisely, that for any quantity **X**, there is some **Y** that is correlated with **X** but is probabilistically independent of other causes of **X**.[8]

I shall leave the proof as an exercise, but for an initial intuitive motivation, note that, when **A** causes **B**, any quantity correlated with **A** will either be a cause or effect of **A** or share a common cause with **A**, and so will be a cause of **B** or share a common cause with **B**, and so by Reichenbach's Principle will be correlated with **B**; but nothing requires a cause of **B**, or something that shares a common cause with **B**, to be correlated with **A**.

8 Counter-examples to Reichenbach's principle

Probabilistic theories of causation are no good. Since they equate causal structure with the correlational patterns that manifest them, they are committed to viewing the bridge principles as metaphysically necessary truths. However, far from being necessary, the bridge principles are not even true. There are counter-examples to both Reichenbach's Principle and the Faithfulness Condition. I shall discuss the former in this section and the latter in the next. Many of the issues will be familiar, so I shall endeavour to be brief.

Three kinds of correlations are prima facie counterexamples to Reichenbach's Principle: (1) times series correlations, (2) collider bias correlations, (3) EPR correlations. In all these cases we would seem to have genuine correlations between quantities even though, contrary to Reichenbach's Principle, they are not causally linked—neither is causing the other nor do they have a common cause. Let me first explain these cases, and then consider what defenders of the probabilistic project might say in response.

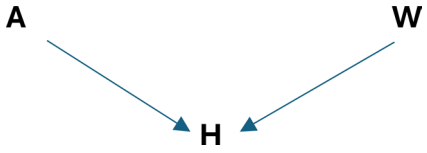
(1) Time Series Correlations. These arise when two independent times series are both monotonically increasing or decreasing. The classic example, due to Sober (2001), involves Venice water levels and London bread prices since 1800. Both have been steadily rising, and this alone means they will display a positive correlation: knowing the value of one is itself strongly informative about the value of the other.

about failures of faithfulness.

[8] It is worth noting that this analysis does need any explicit requirement of causal sufficiency. This is because its assumptions ensure that any **A**-**B** correlation due to their having a common cause will be distinguished from correlations due to one of **A** or **B** causing the other by the fact that *both* **A** and **B** have independent sources of variation.

Yet we certainly don't want to say that because of this the water levels must be causing the bread prices or vice versa, or that they must have a common cause.

(2) Collider Bias Correlations. These arise when we control for the common effect of two joint causes in a "collider" structure. Take my earlier example of anxiety, weight and hypertension:



In a case like this, we can expect that among people with a given level of hypertension, anxiety **A** will be *negatively correlated* with weight **W**. (To see why, note that high (low) hypertension **H** requires high (low) anxiety **A**, or high (low) weight **W**, or both. So knowing that someone with high (low) hypertension **H** has high(low) anxiety **A** contra-indicates their having high (low) weight **W**, and vice versa.)

Now imagine that our background field includes a given level of hypertension **H**. That is, the population of interest is restricted to people with that value of **H**. Within this population, **A** and W will be negatively correlated. Yet by hypothesis they are not causally linked—neither is causing the other nor do they have a common cause. This then gives us another counterexample to Reichenbach's Principle.

(3) EPR Correlations. For this third kind of counterexample, it will be helpful to focus on the "Principle of the Common Cause". This principle—which is an immediate corollary of Reichenbach's Principle—asserts that:

> If **A** and **B** are correlated, then one must be causing the other or they must have one or more common causes, and in the latter case controlling for the common causes will screen off the correlation.

This principle seems to be violated by the quantum-mechanical correlations displayed by results of measurements on spacelike separated entangled particles ("EPR" correlations henceforth, after Einstein et al., 1935). A natural first thought is to view these correlations as an upshot of the particles' states having a common cause, namely the state of the source that emitted them. However, if that were right, then according to the Principle of the Common Cause these correlations ought to be screened off from each other by that state. However, as John Bell showed with his eponymous inequality, orthodox quantum theory implies that EPR correlations will display a collective structure that mathematically rules out their being screened off by any possible state of their source (Bell, 1964)—and in the early 1980s a series of experiments by Alain Aspect showed that quantum theory is empirically correct in this respect (Aspect et al., 1981).

In truth, it is not clear how to interpret EPR correlations causally. Different interpretations of quantum mechanics have different implications about the causal structure involved. Still, for our purposes we need only observe that EPR correlations certainly pose a challenge to Reichenbach's Principle. After all, they are genuine correlations, by anybody's counting. So, by Reichenbach's Principle, either one mea-

surement is causing the other, or they have a common cause. But in the former case special relativity would require the two measurements not to be spacelike separated, which they aren't, and in the latter case Reichenbach's Principle says that the common cause should screen off the correlation, which Bell's inequality tells us it can't. So, whichever way we turn it, Reichenbach's Principle is in trouble.

In response to these three challenges to Reichenbach's Principle, defenders of the probabilistic project standardly argue that they involve the *wrong sort of correlations*. That is, their strategy is to be more specific about the kind of correlations that they take to ground causal facts. Not any correlations can play this role. We need to exclude correlations due to time series, or collider bias, or measurements of entangled quantum systems.

The challenge for the probabilistic project is to explain what is wrong with these correlations without invoking causal notions. If we are aiming for a metaphysical reduction of causation to correlations, we cannot legitimately appeal to causal facts in specifying the grounds of the reduction.

It is an interesting question whether this challenge can be met. Maybe the times series correlations can be dismissed on the grounds that they illegitimately generalise over spatiotemporally related objects. And perhaps the collider bias correlations can be dealt with along the lines used earlier to finesse the need to appeal to causal sufficiency—that is, they will arguably be exposed as bogus once we expand our study to include the collider as a quantity in its own right.[9] As for the EPR correlations, it is hard to decide whether they can reasonably be accommodated without taking a stance on the interpretation of quantum mechanics.

In any case, I shall not pursue these issues any further at this stage. Even if the probabilistic project can somehow accommodate these counterexamples to Reichenbach's Principle, I am about to offer some more compelling objections. These will lead us to a deeper understanding of causal inferences and the bridge principles they rest on. Once we have this we will be better placed to comment on the counterexamples to Reichenbach's Principle.

9 Failures of faithfulness

Let me now turn to the Faithfulness Condition. This too is open to counterexamples, and these are more directly damaging for the probabilistic project than the counterexamples to Reichenbach's Principle. This is because they don't invoke unusual species of correlations, but simply depend on the way that ordinary causal structures can conspire to not to reveal themselves in correlations.

These are "failures of faithfulness". They occur when one quantity causes another via two different paths, with the positive influence on one path cancelling out the negative influence on the other, resulting in a null correlation between the cause and effect. The classic example, due to Hesslow (1976), supposes that the direct positive influence of birth control pills **B** on thromboses **T** is precisely cancelled out by

[9] I would like to thank one of the readers for *Synthese* for suggesting this option to me.

its negative influence, due to blocking pregnancies **P** which themselves conduce to thromboses, as in this causal structure:

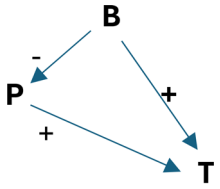
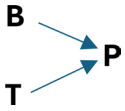


Diagram 4

If the two paths of influences from **B** to **T** are precisely balanced, then the overall result will be that thromboses are no more common among women who take birth control pills than among those who don't. So that would be a counterexample to the Faithfulness Condition's claim that absence of correlation implies absence of any causal link. Here **B** and **T** are not correlated, but they are causally linked all right—indeed linked twice over, both because of **B**'s direct influence on **T** and because if its indirect influence though **P**.

Such failures of faithfulness[10] present a direct challenge to the probabilistic project. To drive the point home, note that in Hesslow's example birth control pills **B** and thromboses **T** are overall unconditionally uncorrelated, but both are correlated with pregnancy **P**—which according to the bridge principles unequivocally implies this fallacious causal structure instead of the real set-up:



Now, it is true that such perfect cancelling out would be an unlucky freak. And this perhaps argues that we can dismiss the possibility when we are engaged with the practical business of inferring causes from correlations in real life.[11] But this dismissal is not acceptable if we are aiming at a metaphysical reduction of causation of the kind essayed by the probabilistic project. For this account says that causal structure is nothing but correlational structure, and so it needs to hold that cases where they come apart are not just unlikely, but downright metaphysically impossible. And the trouble is that cases like Hesslow's do not seem at all metaphysically impossible, however unlikely they may be.

[10] A simpler example of a faithfulness failure is due to Pearl (2000 48). A bell rings if two fair coins show the same face. The outcome of each coin toss is a cause of the bell ringing, but neither is correlated with it. This example too hinges on a sort of happenstantial cancelling-out. The dependence of the ringing on the coin tosses is precisely structured so that at the level of correlations the influence of one coin masks the influence of the other.

[11] A nearby danger, however, is a real issue for empirical researchers. Even if exact cancelling of underlying lawlike correlations would be a freaky coincidence, approximate cancelling is all too likely to mislead researchers who have no alternative but to estimate population independencies from finite sample statistics.

10 Single-case causation

Failures of faithfulness make it clear that there is more to causal structure than is revealed by correlational patterns. The causal structures themselves lie deeper than the correlations that provide input to the bridge principles, and this allows the causal structures to conspire to conceal themselves from the techniques of causal inference.

This moral can be confirmed by some simple observations about the relation between *single-case* causal claims and the kind of causal generalizations that issue from causal inference, like *smoking causes cancer*.

I take it that the bridge principles in conjunction with actual empirical correlations have satisfactorily established that smoking causes lung cancer. Now suppose that Jane Doe smokes and gets lung cancer. Did her smoking cause her lung cancer? It depends. Even if it is true that "smoking causes lung cancer", in the sense inferred from the empirical correlations courtesy of the bridge principles, Jane's strong immune system might prevent cigarettes from harming her and she might have acquired her lung cancer from asbestos exposure instead.

This again shows that there must be more structure to causation than is captured by the kind of causal claims that issue from causal inference. At best generic claims like *smoking causes lung cancer* only give us partial information about the causal links between smoking and cancer. Even when they are true, they do not tell us about all the different factors that can exert a causal influence on the effect at issue, and so do not give us enough information to assess causal claims about particular cases.

A full analysis of causation needs to show us how to uncover this extra structure. I shall now propose such an account.

11 Recursive laws with exogenous independence

The traditional way for medical scientists to analyse our weight-anxiety-hypertension example would be to posit a set of regression equations like this:

$$\mathbf{W} = \mathrm{e}_{\mathbf{W}}$$

$$\mathbf{A} = \mathrm{e}_{\mathbf{A}}$$

$$\mathbf{H} = \mathrm{a}\mathbf{W} + \mathrm{b}\mathbf{A} + \mathrm{e}_{\mathbf{H}}$$

In this kind of regression analysis we start with a set of macroscopic quantities—**W**, **A**, **H**. The subscripted rightmost e-quantities are then exogenous "noise terms" that include enough microscopic fine-grained detail to make the equations deterministic. The equations are *recursive* in that they can be ordered so that no macroscopic term appears as an independent term on the right-hand side of an equation unless it has already appeared as a dependent term on a left-hand side.

These equations are deterministic, but at the same time the quantities involved in regression analysis are to be understood as random variables with a joint probability

distribution. Equations like these thus have both a probabilistic and a deterministic aspect. It is this combination that will hold the key to understanding causal inference.

Now, it is natural to read regression equations *causally*, taking dependence in the equations to represent causal dependence. So in the above equations **H** is naturally viewed an effect of both **W** and **A**.

But of course this reading cannot be grounded in the content of the equations alone. After all, equations can be reordered and still impose the same joint constraints on the quantities they involve. For example, the following equations are equivalent to those we started with.

$$\mathbf{H} = \mathrm{e}^*_{\mathbf{H}}$$

$$\mathbf{W} = \mathrm{c}\mathbf{H} + \mathrm{e}^*_{\mathbf{W}}$$

$$\mathbf{A} = \mathrm{d}\mathbf{H} + \mathrm{e}^*_{\mathbf{A}}$$

where $e^*_{\mathbf{H}} = ae_{\mathbf{W}} + be_{\mathbf{A}} + e_{\mathbf{H}}$; $e^*_{\mathbf{W}} = -b/a\ e_{\mathbf{A}} - e_{\mathbf{H}}/a$; $e^*_{\mathbf{A}} = -a/b\ e_{\mathbf{W}} - e_{\mathbf{H}}/b$; $c = 1/a$; $d = 1/b$.

However, this rewriting now presents a different order of dependency among the macroscopic quantities, with **A** and **W** depending causally on **H**, rather than vice versa.

Still, while these two sets of equations impose the same joint constraints on the quantities, they are not the same in all respects. In particular, if the original exogenous noise terms were probabilistically independent of each other, then the starred noise terms will not so be probabilistically independent, as inspection immediately shows.

Recall an idea we met from Sect. 4 above—every quantity has some source of variation that is independent of its other causes. In the present context, we can use this idea to tell us when an ordering of equations correctly represents causal order—it will do so only if the ordering renders the exogenous noise terms probabilistically independent.

I have illustrated the idea of recursive equations with linear regression analysis. But recursive equations need not be restricted to linear dependencies between real-valued quantities.

For a start, we need not restrict ourselves to real-valued quantities. We can also work with binary properties, determinable properties, ordinal rankings, rational quantities and so on. And these need not be linked just by linear equations, but by any system of laws specifying some pattern of co-variation between values of these quantities.

So from now on I shall be interested in cases where, starting with a set of macroscopic quantities $\mathbf{X}_1, \ldots \mathbf{X}_n$, we can arrange them, together with a corresponding set of noise terms $e_{\mathbf{X1}}, \ldots e_{\mathbf{Xn}}$, into an ordered system of deterministic laws of the form

$$\mathbf{X}_i = \mathrm{F}(\mathbf{X}_1, \ldots \mathbf{X}_{i\text{-}1}, e_{\mathbf{Xi}})$$

in which the exogenous noise terms are all probabilistically independent of each other.

12 Causation as RLI ancestry

In such cases I shall say that we have a Recursive structure of deterministic Laws with exogenous Independence—an "RLI" for short. And the thesis I now want to explore is that causation is constituted by such RLIs. More precisely, I hypothesise that:

> **THESIS**: **X** causes **Y** if and only if it is an ancestor of **Y** in a recursive structure of deterministic laws with independent exogenous noise terms.

We might usefully view this as a return to a *regularity* theory of causation. Causal relations are grounded in the way that certain quantities covary with others in line with natural laws[12]. But unlike older regularity theories of causation, this account appeals to laws that have a direction built into them. The quantities in an RLI are ordered, thanks to the requirement that the exogenous noise terms must be probabilistically independent, and so it is always clear which of two causally related quantities is the cause and which the effect.

Now, my Thesis above can be challenged in both directions. Are there really RLIs behind every single instance of causation? And are there not some RLIs whose direction runs counter to causation?

It would take us too far afield to discuss these twin challenges at any length. In the remainder of this paper I shall rather defend my thesis by showing how it offers natural explanations for a wide range of phenomena, including causal inference, single case causation, and the connection between causation and rational action.

Still, before proceeding it will be worth briefly indicating in outline how the two challenges above might be met.

The first challenge asked why an RLI should be available to ground every instance of causation. In this connection it is relevant that *thermodynamic* asymmetries seem to rest on structures that are very similar to RLIs. The standard statistical mechanical analyses of the behaviour of non-equilibrium thermodynamic systems also invoke deterministic equations with probabilistically independent microscopic initial conditions (Wallace, 2023, forthcoming). The ubiquity of asymmetric thermodynamic processes thus gives us reason to suppose that structures of laws with exogenous independence are pervasive throughout nature.

To digress for a moment, this affinity with statistical mechanical accounts of thermodynamic behaviour argues that causation is an essentially macroscopic phenomenon, depending on structures which emerge only with respect to coarse-grained aspects of physical systems. This does not mean that microscopic quantities cannot enter into causal relations. The relevant structures can include microscopic causes and effects alongside macroscopic ones.[13] But it does mean that we will lose sight

[12] I shall stand neutral on the nature of laws. I intend my account of causation to be compatible with any analysis of lawhood, including both best-systems and necessitarian accounts.

[13] While the distinction between macroscopic and microscopic quantities is inevitably vague, perhaps we can follow statistical mechanics in implicitly demarcating non-gerrymandered macroscopic quantities simply as those which display emergent lawlike behaviour.

of causation when we characterise systems solely in fundamental physical terms—which is scarcely surprising, given that microscopic physics is temporally symmetric where causation is not.[14]

The second challenge related to the possibility of RLIs that don't match actual causal relations. Now it is certainly possible in some cases to contrive empirically adequate recursive systems of equations with exogenous probabilistic independence that do not line up with real causal relations. But these constructions all involve various artificialities, and so can be dealt with by tightening up the requirements for RLIs in the Thesis above. As I said, this is not the place to pursue these issues in detail, so for now let me simply state that some of these unwanted RLIs can be dismissed on the grounds they are not genuinely lawlike, others because they cannot be embedded in more inclusive RLIs involving further quantities, and others again because they do not respect the distinction between macroscopic and microscopic quantities. (For an extended discussion of such counterexamples, see Papineau forthcoming Chap. 6.)

Rather than further elaborating my responses to these two challenges here, I shall now aim to defend my Thesis by showing how nicely many puzzling aspects of causation fall into place once we adopt it. Let me start by showing how the bridge principles follow naturally.

13 Deducing Reichenbach's principle and explaining the faithfulness condition

Suppose we have some set of quantities $\{\mathbf{V}\}$ that are covered by an RLI—a recursive structure of deterministic laws with exogenous probabilistic independence. Then any quantity in $\{\mathbf{V}\}$ will ultimately descend from some set of exogenous noise terms. Consider two such quantities $\mathbf{X}$ and $\mathbf{Y}$. If they descend from disjoint sets of noise terms, then they will be probabilistically independent, because the RLI ancestors that determine them are. So conversely, if they are correlated, then they must share ancestors or one must descend from the other in the RLI. And, more generally, if they are correlated conditional on the quantities in some $\{\mathbf{Z}\}$, then they must be linked in the RLI in ways that don't involve $\{\mathbf{Z}\}$, because without such non-$\{\mathbf{Z}\}$ linkage their residual variation when $\{\mathbf{Z}\}$ is held constant would again be probabilistically independent.

So far this is just a mathematical result (Pearl, 2000 Theorem 1.4.1) which makes no mentions of causation. It simply tells us how a correlation between two quantities in an RLI is inconsistent with disjoint provenance. But if we now conjoin this result with my Thesis, then Reichenbach's Principle is assured for any quantities that are covered by RLIs—not only do correlations tell us about RLI linkage, but also about *causal* linkage.

[14] Some might object that they can conceive of a causal world without macroscopic structure. Consider, for example, a world with nothing else in it where one perfectly hard particle bumps into another and causes it to move (Cf Ehring, 1987, Sosa & Tooley, 1993 Introduction). My response is that we might be able to conceive of such a world, but we would be conceiving a metaphysical impossibility (Papineau, 1988).

Note, however, that this result only applies to quantities that *are* covered by an RLI. This now casts a new light on the counterexamples to Reichenbach's Principle discussed in Sect. 8 above—the time series, EPR and collider bias correlations. These are genuine correlations all right, but we can now say that the reason they do not evidence causation is that they are not generated by an underlying RLI. They are manifestations of different kinds of structures, and so should not be read as indicating any causal linkage.

Take first the time series correlations. We can perfectly well write down time series equations for the London bread prices, with the price each year depending on the price the previous year plus some exogenous error term, and similarly for the Venice water levels. And if we assume the error terms are all probabilistically independent, this will indeed account for the way that the bread prices and the water levels for any given year will be correlated.

Note, however, that this correlation is between two features of *years*. This means that an RLI-type explanation of this correlation needs would need to generalise over single years and specify lawful connections between their features. But that is not what the time series equations do. Rather they tell us how the London bread prices in one year relate to those *in the previous year*, and the same for the Venice water levels. This just isn't the kind of RLI equational structure to which I am reducing causal relations. That is why time-series correlations do not constitute evidence for a causal link between the quantities they relate.

A similar point applies to collider bias correlations. These arise, remember, when some **A** and **B** are unlinked joint causes of some common collider effect **E**, and we focus specifically on the probability distribution displayed when some value of **E** is held fixed. Within this distribution we will find an unconditional negative correlation between the unlinked **A** and **B**, thus posing a threat to Reichenbach's Principle.

From my present perspective, however, I can simply say that Reichenbach's Principle shouldn't be expected to hold in this case—again because the probability distribution at issue has not been generated by an RLI. To see why, suppose that a genuine RLI relates **A**, **B** and **E**. If we now restrict this RLI to a fixed value for **E**, so to speak, the exogenous independence will be destroyed. The equations will stay the same, but the exogenous terms for **A** and **B** can't still be probabilistically independent, given that **A** and **B** are now correlated even though unlinked in the equations. So once more the equations behind the worrisome correlation don't constitute the kind of structure to which I am reducing causation, and the correlation is thus no evidence for causation.

As for the EPR correlations, they don't come from RLIs simply because RLIs are deterministic and the outcomes in question are not determined. I shall come back to EPR correlations in Sect. 16 where I discuss issues raised by indeterminism.

Note that Reichenbach's Principle now needs to be qualified. Instead of stating it in full generality, as I did earlier, as licensing inferences from any correlations whatsoever to causal conclusions, I shall understand it henceforth as applying only to correlations between quantities that are governed by RLIs.

Let me now turn to the Faithfulness Condition. Unlike Reichenbach's Principle, this does not follow from my Thesis even for quantities that are governed by RLIs.

Rather my Thesis implies that the Faithfulness Condition will generally hold for such quantities, but not exceptionlessly. That is just as it should be.

As a rule we can generally expect quantities that are linked in a system of equations with exogenous independence to be correlated. The equational links plus the background independence will generally lead to the quantities co-varying. In certain special cases, though, like Hesslow's pregnancy-thrombosis example, a specific cancelling-out of coefficients will mean that equationally linked variables display no overall correlation.

From my present perspective, however, this is now no problem. I am no longer trying to ground causal structures in correlations alone. My Thesis is rather that causation is constituted by underlying structures of laws with exogenous independence, not by the surface correlations issuing from those structures. It is no objection to this Thesis that the coefficients in such structures should sometimes conspire to make the correlations misleading about the underlying causal structure.

So the Faithfulness Condition now falls into its rightful place, as something that empirical researchers can generally rely on, but is in principle open to exceptions. There is nothing in the nature of causation to guarantee that probabilistic independencies should not arise by a cancelling out of parameters. This would be a surprising fluke, but it is not built into the nature of causation.

Let me finish this section with a general comment. It is striking how radically non-parametric causal inference now appears. By equating causation with ancestry in RLIs, I have provided a rationale for Reichenbach's Principle and the Faithfulness Condition. However, my rationale did not need to assume anything about the form of the relevant RLIs, nor about the coefficients in their laws. The assumption of deterministic laws and exogenous probabilistic independence alone was enough to justify drawing causal conclusions from nothing but qualitative facts about which quantities are and aren't correlated with which.

Of course, from a practical point of view, it will often be important also to know *how much* one quantity causes another, and whether it *interacts* with other quantities in doing so, and so on. But even so it is remarkable that correlations and independencies alone can inform us that one quantity *does* cause another even while leaving us quite in the dark about such further matters.

14 Causal models and single-case causation

Let me now consider single-case causation. In fact, as will be familiar from the literature on David Lewis's counterfactual theory of causation, there are two single-case relations that we need to consider: counterfactual dependence and "actual" causation.

One fact **E** counterfactually depends on another distinct fact **C** just in case **E** would haven't occurred if **C** hadn't. This relationship is sufficient for **C** actually causing **E**, but it is not necessary. For example, Jane Doe's lung cancer could actually be caused by her smoking even if it doesn't counterfactually depend on it. Imagine a case where Jane was exposed to asbestos as well as being a smoker, and this asbestos exposure would have given her lung cancer, if only the smoking hadn't done so first. In that case, even though Jane's smoking actually caused her lung cancer (now I'm suppos-

ing her immune system didn't counteract the smoking), it's not true that she wouldn't have contracted lung cancer if she hadn't smoked.

Much recent work on both counterfactuals and actual causation appeals to "causal models" (Hitchcock, 2018). A causal model is a set of deterministic laws specifying how certain quantities asymmetrically depend on others. (So they are exactly like my RLIs, save that the asymmetries are normally assumed rather than explained—I'll come back to this in a moment.)

Once we are given a causal model for a particular situation, there is a straightforward recipe for evaluating counterfactuals. We just hold all the exogenous quantities on which the antecedent does not depend fixed at their actual values, switch the antecedent itself to its counterfactual value, crank the equations, and see if the consequent follows (Galles & Pearl, 1998; Briggs, 2012).

Actual causation is trickier. The basic strategy for deciding whether **C** actually caused **E** is to see whether the latter counterfactually depends on the former when "off-path" quantities are held fixed at their actual values. For example, did Jane's cancer counterfactually depend on her smoking once we hold fixed the absence of some of the later steps in the process by which asbestos causes cancer? However, some cases seem to demand a more complicated treatment in which certain off-path quantities are fixed at values other than their actual ones (Weslake, 2013). This is not the place to pursue these intricacies.

However the details pan out, I am entirely in favour of analysing counterfactual dependency and actual causation in terms of causal models. Not that causal models solve all the problems. Apart from the complexities attaching to actual causation, there are also questions about how to choose causal models to start with. The choice of a model for some specific single-case situation is in effect a commitment to regard quantities not included in the model as part of a fixed background field. Such choices can make a difference to the evaluation of both counterfactuals and actual causation, yet it is a matter of debate how far they answer to anything beyond pragmatic considerations. Still, I do not regard this as any reason to turn away from causal models in analysing single-case relations. In my view, these models point us in the right direction, even if they leave further questions open.

The more fundamental issue is the meaning of causal models, and in particular of the asymmetric dependence relations they involve. The literature is curiously reticent on this matter (Menzies & Beebee, 2024 Sect. 5.3). After all, it doesn't seem that much of a challenge to analyse asymmetric single-case relations if we are allowed to start with asymmetric deterministic type-dependencies between quantities. The more substantial undertaking is to explain where the asymmetries come from in the first place. Yet most of the literature seems happy to use the asymmetric models without explaining them.

It take it to be a major virtue of my account of causation that it fills this gap. My RLIs are perfectly suited to explaining the asymmetry of causal models. RLIs are constituted by systems of laws detailing how certain quantities are asymmetrically determined by others, with the asymmetry imposed by the requirement that the exogenous quantities be probabilistically independent.

15 Rational action

As we have seen, the generic conclusions delivered by causal inference—like *smoking causes cancer*—are on their own too coarse-grained to determine single-case causal facts—was Jane Doe's cancer counterfactually dependent on, or actually caused by, her smoking? To decide single-case facts, we need to know about all the factors relevant to the effect in question and how they apply to the particular case at hand—which we generally don't.

This might make us wonder about the point of causal inference. Why are we so interested in its generic conclusions, if they are going to leave us in the dark about single cases?

The answer is that these conclusions are important for a quite different reason. They can provide agents with a guide to rational action, even when those agents don't know about the precise details of their situation and so about the single-case facts. The generic claim that *smoking causes cancer* can inform Jane Doe that it's not a good idea to smoke, even while she remains ignorant as to whether her smoking would indeed make the difference between getting cancer and not.

This works because a generic claim of the form **A** causes **R** tells us that there are *some* circumstances in which **A** will be causally decisive for **R**, in the sense that in them **A** will make up a determining conjunction for **R** when one wouldn't otherwise obtain, and that to this extent we have some reason to do **A** if we are concerned about **R**.

To see why the generic claim tells us this, note that it tells us that **A** is an ancestor of **R** in a structure of directed laws—an RLI. And this means that **A** is one of the independent quantities in a directed law which determines one of the independent quantities in a directed law which determines **R**. And this means that **A**, together with definite values for the other independent quantities in the laws just mentioned, will constitute a determining conjunction for **R**. And in circumstances where those values of those other independent quantities obtain, and no other determining conjunction for **R** obtains, **A** will be causally decisive for **R**.

Of course, it is not much of a guide to action to know that **A** *can* so be causally decisive for **R**. Agents wondering whether to do **A** to affect **R** will also need to know how *likely* it is that **A** will deliver such a determining conjunction when one wouldn't otherwise obtain. How likely is it that my smoking will be decisive for my contracting cancer? Fortunately, causal inference will often deliver this kind of information too, by telling us about the "average effect" of **A** on **R**.

To see how this works, let us first suppose that **A** is probabilistically independent of any other causes of **R**. In this simple case, the average effect of **A** on **R** will be given by the difference between Prob(**R**|**A**) and Prob(**R**|not-**A**). This difference will tell us precisely the probability of **A** being causally decisive for **R**.[15]

[15] A little proof. Let **X** be the disjunction of factors together with which **A** forms a determining conjunction for **R**, and let **Y** be the disjunction of other determining conjunctions for **R**. Then, in the presence of **A**, **R** will occur if and only if **X** or **Y**, whereas in the absence of **A** it will occur if and only if **Y**. The difference between Prob(**R**|**A**) and Prob(**R**|not-**A**) thus reflects the probability of **X** obtaining, and thus of **A**'s being causally decisive.

More precisely:

In the more general case, where a potential action **A** is correlated with other causes **O** of the result **R**, we need something a bit more complicated. Here the simple difference Prob(**R**|**A**) - Prob(**R**|not-**A**) will no longer measure the average effect of **A** on **R** but will be inflated (or deflated) to the extent that **A** is positively (or negatively) associated with other positive (or negative) causes **O** of **R**. So now an agent needs to know how likely it is that **A** will be causally decisive for **R** *apart from* its association with any other causes of **R**.

The way to measure this is to look at the probabilistic association between **A** and **R** for fixed values of the other causes **O** of **R**. Thus Prob(**R**|**A** & **O**) - Prob(**R**|not-**A** & **O**) will measure how often **A** will be causally decisive for **R** when **O** obtains. And similarly Prob(**R**|**A** & not-**O**) - Prob(**R**|not-**A** & not-**O**) will measure how often **A** will be causally decisive for **R** when not-**O** obtains. To work out the overall average effect of **A**, then, we need to take a weighted average of the different differences **A** might make to **R** for different values of **O**, weighed by the probability of those different values of **O**. This figure will then tell an agent ignorant of the relevant details of their situation how likely it is that doing **A** will be causally decisive for **R**.

16 Indeterminism

On most interpretations, quantum mechanics gives us reason to suppose that our world is indeterministic.[16] This makes a number of complications for the analysis so far.

For a start, indeterminism complicates the explanation of causal inference in terms of recursive structures of laws with exogenous independence – RLIs. As I explained above, the bridge principles assumed in causal inference follow from the exogenous independence and determinism of RLIs. But the determinism was crucial here. Without it the bridge principles do not follow.

To illustrate, let us focus once more on the Principle of the Common Cause. This says that correlations due to one or more common causes will be screened off if we control for those common causes. Along with the more general Reichenbach's Principle, this follows from my Thesis equating generic causation with RLI-ancestry.

We are supposing **R** ←→ **AX** or **Y**.

So Prob(**R**|not-**A**).

= Prob(**Y**).

given that **Y** is probabilistically independent of **A.**

And Prob(**R**|**A**).

= Prob(**AX** or **Y**|**A**) = Prob(**Y**) + Prob(**X**) - Prob(**XY**).

given that **Y**, **X**, and **XY** are probabilistically independent of **A.**

So Prob(**R**|**A**) - Prob(**R**|not-**A**).

= Prob(**X**) – Prob(**XY**).

which is the probability A will form a determining conjunction (with **X**), less the probability that determining conjunction will in any case obtain (because **Y** is also present).

[16] I have in mind here stochastic interpretations like GRW and also Everettianism. (While the latter takes the multiverse to be deterministic, it implies effective indeterminism within any "branch" of reality.)

Once we control for common causes, any remaining variation in joint effects will be due to disjoint sets of probabilistically independent exogenous terms, which will then force the joint effects to be conditionally uncorrelated. However, if the dependent quantities here are not *determined* by their RLI-ancestors. then this conditional probabilistic independence does not follow, since there will then be logical room for those quantities to coordinate themselves, so to speak, even if those ancestors are probabilistically independent.

As it happens, this is not just a logical possibility. The EPR quantum mechanical correlations are a real-world realization of just this phenomenon. After all, in an EPR set-up, the antecedents of the spatially separated measurement results are the particles' common source plus probabilistically independent influences on the measuring instrument settings. If the measurement results were *determined* by these antecedents, then any correlation between the measurement results would disappear conditional on the state of the source, because of the probabilistic independence of the instrument settings. But as we have seen the EPR measurement results are correlated in ways that mathematically preclude such screening off by any possible feature of the source.

Does this mean that the principles of causal inference fail to apply in the real indeterministic world? Not really. The kind of coordinated results involved in EPR correlations only manifest themselves under very carefully controlled conditions. The wave function of the paired particles must not "collapse" until the coordinated measurements have been performed. In the ordinary way of things, however, entangled particles will interact with any manner of macroscopic objects before such measurements, none of which will be designed to display their entanglement. So the results of such interactions will generally be probabilistically independent, with the consequence that the causal inference principles will still work as a guide to causal structure. Even in the real indeterministic world, the laws governing events will still give us "pseudo-deterministic" RLIs, so to speak, in the sense that they will display the same correlational signatures of their structure as if they were deterministic, and as a result will conform to the bridge principles.

Indeterminism is also relevant to causation in further ways. The analysis of single-case causation I offered earlier in terms of "causal models" (that is, RLIs) assumed determinism. Conclusions about counterfactual dependence, and about actual causation, were drawn from the way certain causally independent quantities did or did not *determine* causally dependent quantities. What then should we say in cases where the relation between causally independent and dependent quantities is pseudo-deterministic rather than deterministic? The existing literature is divided on this question. Some allow that a result can be actually caused by another, or counterfactually dependent on it, even when it is undetermined. Others say that in such cases we should only speak of the causation or dependence of *chances* rather than actual results. Others doubt that the relevant claims have truth conditions at all, as opposed to degrees of assertibility. I am not sure that there are facts of the matter here. There's no reason why our everyday causal notions should be able to adjudicate all the hard questions thrown up by indeterministic metaphysics. Once we are clear about the substantial issues, perhaps all that remains is to stipulate a convenient way to talk.

More substantial issues arise in connection with the way probabilistic causal claims provide guides to action. In the previous section I explained how the different probabilities for some **R**, conditional on values for an action **A** and **R**'s other causes, can be taken to indicate the probability of **A** contributing to a conjunction that counterfactually determines **R**. The basic idea was that these conditional probabilities tell us how likely it is on balance that **A** will ensure a determining conjunction for **R** when such a determining conjunction wouldn't otherwise obtain—with agents concerned about **R** thus having that much of a reason to do **A**.

An underlying assumption here is that rational agents will typically have less than full information about their situation. Even when we suppose that the actual features of their situation, plus their actions, will *determine* whether or not the outcome of interest occurs, they will not generally know about all those features of their situation, and so will need to act on the basis of how *probable* it is that alternative actions, in conjunction with those features, will determine the outcome.

The same general picture applies even if the world is indeterministic. Now the features of actual situations, plus agents' actions, will only fix *chances* for the outcomes of interest, rather than determining whether they occur. Still, the same general points apply. As before, typical agents will not know about all the relevant features of their situation, and so will still need to choose on the basis of the *average effect* of their alternative actions on the outcome. And this average effect will still be given by the conditional probability of the outcome given the relevant action and the outcome's other causes. However, now this average effect will measure, not the probability that a given action will causally *determine* the outcome, but the weighted average of the different *chances* the action might cause for the outcome, weighted by the probabilities of the different situations the agent might be in.

So from a practical point of view it turns out indeterminism makes no difference to rational agents. Whether or not determinism is true, rational agents still need to make decisions on the basis of the causal difference their actions make to the probability of desired outcomes.

17 Agency and causation

In these last two sections, I have implicitly been assuming that *causation* is prior to *rational action*. I have been taking it as given that agents have reason to do **A** in pursuit of **R** to the extent the chance of **R** *causally* depends on **A**. Many philosophers of causation hold that this has things the wrong way round. In their view, rational action is metaphysically prior to causation. We should be aiming to explain causation in terms of rational action, not vice versa.

One version of this idea is suggested by occasional comments Judea Pearl, James Woodward and others. Their basic idea is that the species of co-variation between actions and their desired results is very different from that obtaining between ordinary pairs of observed events, and that causation can be equated with this species of co-variation.

According to this idea, when we are dealing with ordinary pairs of observed events, we can capture their mode of co-variation via straightforward unconditional correla-

tions. But we need something different when we are dealing with the co-variation of *actions* and *desired results*. We don't want to predict the results of our actions using possibly spurious unconditional correlations. Rather we should anticipate results in line with the conditional correlation between action **A** and result **R** *conditional* on other causes of **R**, that is, the correlation that remains when the influence of those other factors is filtered out. And this distinctive mode of co-variation, so the thought goes, is what constitutes causation.[17]

However, there is no real substance to the idea of a distinctive species of co-variation involving actions and results. As far as co-variation itself goes, unconditional correlations are all we need. After all, if we really want to *predict* what will happen after some action has occurred, simple unconditional correlations are perfectly adequate to the task. (Smoking will still be a perfectly good *predictor* of cancer, even if both are independent effects of a genetic factor and smoking doesn't *cause* cancer at all.)

What unconditional correlations are not good for, though, is *guiding rational choices*, precisely because rational choices need to have a *causal* impact on the chance of desired results, rather than just be correlated with them. And to capture this causal impact, we need the relevant conditional correlations, not unconditional ones, for the reasons laid out earlier.

This now shows that the Pearl-Woodward line of thought under examination does not lead us to an alternative to the causal analysis of rational action I have offered, but rather a version of it. First we need to identify causal relationships, and then we can say that rational actions are those that causally influence desired results.

18 Interventions

However, this basic point can be, and often is, confused by talk of "interventions".

Let us agree that rational choices should be guided by the correlation between the putative action and the desired result *conditional* on all the other causes of that result—the "average effect" of the action on the cause. Now, note that only those other causes that are themselves correlated with the putative action are able to make

[17] A particularly clear expression of this idea can be found in Christopher Hitchcock's *Stanford Encyclopedia of Philosophy* entry on "Causal Models" (2018):

"A conditional probability such as P(Y=y|X=x) gives us the probability that *Y* will take the value *y*, given that *X* has been observed to take the value *x*… However, when we intervene, we override the normal causal structure, forcing a variable to take a value it might not have taken if the system were left alone."

Similarly, in *Causality* (2000) Judea Pearl offers an example in which the slipperiness of the pavement is affected both by the season of the year and by whether a sprinkler has been turned on:

"… whatever relationship existed between seasons and sprinklers prior to the action, that relationship is no longer in effect while we perform the action. Once we physically tum the sprinkler on and keep it on, a new mechanism (in which the season has no say) determines the state of the sprinkler" 41.

Again, in his *Making Things Happen* (2003), James Woodward argues that, if we start with a spurious observed correlation between some putative cause **C** and some putative effect **E**, and then look to see what happens in those cases where some agent seeks to actively produce **C** in pursuit of **E**, we will find that:

"… the correlation between **C** and **E** will not be stable or invariant under efforts to use **C** to control **E**" 32.

the action seem more (or less) causally effective that it actually is. This then implies that, *if* some action *were* uncorrelated from all the other causes of its target result, without disrupting other dependencies between quantities, *then* its simple unconditional correlation would be the right measure of its causal influence on the result.

Suppose now that we specify that some putative cause **A** of **R** is an "intervention" if it is produced in some way that decorrelates it from any other prior causes of **R**, without disrupting any other dependencies between quantities. Then it would be correct to say that the causal influence of **A** on **R** will correspond to the simple unconditional correlation between **A** and **R** that *would* obtain if **A** *were* an intervention.

This correspondence of course provides a rationale for the practice of investigating causation by means of *randomized experiments*. In such an experiment, subjects are assigned different values of some putative cause **C** of **E** at random. To this extent[18], **C** will be produced by an "intervention" that forcibly decorrelates it from any other causes of **E**, and thus ensures that any correlation between **C** and **E** observed within the experiment reflects the genuine causal influence of **C** on **E**, unconfounded any influence exerted by other causes of **E** with which **C** is associated outside the experiment.

However, none of these connections between causation and "interventions" amounts to a way of analysing causation in terms of the co-variation of *actions* with results. The notion of an "intervention" is just a technical way of articulating the point that the causal influence of one quantity on another is given by their correlation conditional on the other causes of the putative effect.

A crucial point here is that *actions* are not generally interventions in the technical sense just defined. If they were, then their causal influence on results would indeed correspond to their unconditional correlation with them. But there is no reason why they should be.

For example, suppose level of school funding **S** and household wealth **H** are both causes of examination results **E**—children's educational achievement is affected both by schools and home backgrounds. Now consider a parent deciding whether to send their child to an expensive school. Of course their choice will be influenced by their household wealth **H**—you're more likely to choose an expensive school if you can afford it. But this now means that the parent's choice of school **S** is not an "intervention" with respect to examination results **E**. It is correlated with another cause of **E**, namely household wealth **H**. And, in line with this, the unconditional correlation of schools **S** with examination results **E** will not be a good measure of the causal impact of **S** on **E**. It will be spuriously inflated by the association of **S** with the confounding cause household wealth **H**.

This is by no means a fanciful example. There are any number of ways, both known and unknown, in which background factors will independently influence both choices and desired results, and so stop actions being "interventions". If this were not so, we wouldn't need to worry about "spurious" correlations in estimating

[18] Importantly, the randomization of the cause is just one part of the requirements for an intervention. In addition, the intervention must not disrupt any other dependencies between quantities (Woodward, 2003 98−9). It is not to be taken for granted that randomized experiments satisfy this further requirement. For example, the anxiety occasioned by the randomized assignment of a drug might undermine an effective mechanism that operates in the general population.

causal impact. Retrospective observation of correlations between choices and desired results would suffice for causal inference. But of course the possibility of confounding causes is an ever-present danger—precisely because actions are not generally "interventions".

To sum up this section, it is accurate enough, if not necessarily particularly illuminating, to equate the causal influence of **A** on **R** with the correlation that would obtain if **A** *were* an "intervention". But, since actions are not generally interventions, this doesn't analyse causation in terms of action, but simply repeats the analysis I have given.

19 Woodward's theory of causation

My arguments cast an interesting light on James Woodward's account of causation.

At bottom, Woodward holds that **C** causes **E** if and only if there would be a correlation between **C** and **E** if **C** were produced by an intervention (2003 Chap. 2).

We can usefully view this as akin to the probabilistic theories of causation I discussed in Sect. 6 above. Woodward is explicating causal links in terms of correlations. Of course, as is familiar, Woodward does not offer his account as an explicit reduction, as those probabilistic theories did, because his notion of an "intervention" is itself characterised causally. (We have an "intervention" if the putative cause is produced in a way that decorrelates it from any other *causes* of the result of interest.)

Still, we might well wonder why Woodward is so unambitious. Given that he is prepared to equate causation with correlation under interventions, what stops him following the probabilistic project even further, and saying that causal relations in general are those determined by correlations according to the principles of causal inference? Woodward would still then be able to maintain, if he wished, that **C** causes **E** if and only if they are correlated when **C** is produced by an "intervention"—but, since a full probabilistic analysis will also be able to identify the "other causes" needed to define an "intervention", this dictum need no longer be implicitly circular.

Woodward is prone to insist that correlations need additional causal assumptions to determine causal order (2003 330 342, 2023 Sect. 5). If this were always true, then of course probabilistic theories could not get off the ground. But, as we have seen, there's no reason to suppose it is always true—in all cases where some set of correlations do not suffice for the bridge principles to determine a unique causal order, further possible correlations, perhaps involving further possible quantities, will do so.

Of course, fully embracing the probabilistic project might seem a step backwards for Woodward. After all, as we saw earlier, probabilistic theories are no good. They run into trouble both with correlations that aren't causal—those involving times series, and collider bias, and EPR entanglement—and with causes that fail to show themselves in correlations—faithfulness failures.

But it is not to be assumed that Woodward's own interventionist theory does not face the same problems. Given that he identifies causation with correlations, at least under interventions, we might wonder whether such cases do not cause trouble for his account too.

Let me put the former non-causal correlations to one side. It would take me too far afield to discuss them here, and Woodward himself says nothing about them—though for the record let me say that some of them strike me as potentially problematic for him.

Faithfulness-violating non-correlations raise a more obvious problem, however. As Woodward is aware, an intervention could well instigate two cancelling-out causal paths that leave the cause uncorrelated with the effect. Woodward cites the Hesslow example: intervening to administer birth control pills could fail to alter the probability of thromboses, despite causing them twice over. In response, Woodward seeks to define a notion of a "contributing" cause in interventionist terms and to argue that factors like Hesslow's birth control pills can be "contributing" causes even if they are not "total" causes (2003 Chap. 2). It is a moot point, however, whether this strategy succeeds. In an extended exchange with Woodward, Michael Strevens has argued that Woodward's commitment to explaining everything in terms of interventions means he will always be left with undischarged assumptions about certain interventions not being subject to faithfulness failures (Strevens, 2007, 2008; Woodward, 2008). Strevens' analysis is too intricate to rehearse here—I can only invite interested readers to explore it for themselves. Still, let me conclude by observing that, from my point of view, it would be unsurprising that an attempt to analyse causation in terms of correlations rather than structural equations should founder on this problem.[19]

Data availability As there are no data in this article, this declaration is not applicable.

Declarations

Conflict of interest Not Applicable.

References

Aspect, A., Grangier, P., & Roger, G. (1981). Experimental tests of realistic local theories via bell's theorem. *Physical Review Letters*, *47*, 460–463.

Bell, J. (1964). On the Einstein-Podolsky-Rosen paradox. *Physics*, *1*, 195–200.

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, *160*, 139–166.

Ehring, D. (1987). Papineau on causal asymmetry. *British Journal for the Philosophy of Science*, *38*, 81–87.

Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical description of physical reality be considered complete? *Physical Review*, *47*, 770–780.

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, *3*, 151–182.

[19] I have given talks about this material at the Logic and Metaphysics workshop at CUNY in 2022; the Metropolitan Area Philosophy of Science seminar in New York in 2022; the British Society for the Philosophy of Science conference in Exeter in 2022; the Physics, Time and Agency Conference of the Irish Society for the Philosophy of Time in Dublin in 2022; the European Society for Analytic Philosophy conference in Vienna in 2023; the Time, Chance and Causation Workshop in Stirling in 2024; and the Probabilistic Reasoning in the Sciences conference at the Marche Polytechnic University Ancona in 2024. I would like to thank all those who made comments on those occasions, and in particular Barbara Osimani, Alex Gebharter, Toby Friend, Thomas Blanchard, Franz Huber, and Clark Glymour.

Glymour, C. (2004). Making things happen by James Woodward. *British Journal for the Philosophy of Science*, *55*, 779–790.
Good, I. J. (1961-2). A causal calculus I–II. *British Journal for the Philosophy of Science* 11: 305–318, 12: 43–51.
Hausman, D. (1998). *Causal asymmetries*. Cambridge University Press.
Hesslow, G. (1976). Two notes on the probabilistic approach to causality. *Philosophy of Science*, *43*, 290–292.
Hitchcock, C. (2018). Causal Models in Zalta, E. ed *The Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/archives/sum2024/entries/causal-models/
Menzies, P., & Beebee, H. (2024). Counterfactual Theories of Causation. In E. Zalta (ed), *The Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/archives/sum2024/entries/causation-counterfactual/
Papineau, D. (1988). Response to ehring's 'Papineau on causal asymmetry'. *British Journal for the Philosophy of Science*, *39*, 521–525.
Papineau, D. (1992). Can we reduce causal direction to probabilities? *Philosophy of Science Association*, *2*, 238–252.
Papineau, D. (2001). Metaphysics over methodology: Why infidelity provides no grounds to divorce causes from probabilities. In M. C. Galavotti, P. Suppes, & D. Costantini (Eds.), *Stochastic causality* (pp. 15–38). CSLI.
Papineau, D. forthcoming *Causation* Cambridge: Cambridge University Press.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann.
Pearl, J. (2000). *Causality: Models, Reasoning, and inference*. Cambridge University Press.
Pearl, J. (2018). *The book of why*. Allen Lane.
Reichenbach, H. (1956). In M. Reichenbach (Ed.), *The direction of time*. Dover.
Sober, E. (2001). Venetian sea Levels, British bread Prices, and the principle of the common cause. *The British Journal of Philosophy of Science*, *52*, 331–346.
Sosa, E., & Tooley, M. (1993). Introduction in Sosa, E. and Tooley, M. eds *Causation* Oxford: Oxford University Press 1–32.
Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction and search*. Springer.
Spohn, W. (2001). Bayesian Nets are all there is to causal dependence. In M. C. Galavotti, P. Suppes, & D. Costantini (Eds.), *Stochastic causality*. Stanford: CSLI Publications 157–72.
Strevens, M. (2007). Essay review of Woodward making things happen. *Philosophy and Phenomenological Research*, *74*, 233–249.
Strevens, M. (2008). Comments on Woodward making things happen. *Philosophy and Phenomenological Research*, *77*, 171–192.
Suppes, P. (1970). *A probabilistic theory of causality*. North-Holland Publishing Company.
Wallace, D. (2023). The logic of the past hypothesis. In B. Loewer, B. Weslake, & E. Winsberg (Eds.), *The probability map of the universe: Essays on David albert's time and chance* (pp. 76–109). Harvard University Press.
Wallace, D. Forthcoming probability and irreversibility in modern statistical mechanics: Classical and quantum D. Bedingham, O. Maroney, & C. Timpson (Eds.), *Quantum foundations of statistical mechanics* Oxford: Oxford University Press.
Weslake, B. (2013). A partial theory of actual causation unpublished https://bweslake.s3.amazonaws.com/research/papers/weslake_ac.pdf
Woodward, J. (2003). *Making Things Happen* Oxford.
Woodward, J. (2008). Response to strevens. *Philosophy and Phenomenological Research*, *77*, 193–212.
Woodward, J. (2023). Causation and Manipulability in Zalta, E. *The Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/archives/sum2023/entries/causation-mani/