Why are Causes the Right Way to Produce Effects?—David Papineau
Physics, Time and Agency Conference—Irish Society for the Philosophy of Time
Dublin 28-29 June 2022

A Background

I've elsewhere defended a sort of regularity theory of causation with built-in causal asymmetry. It is motivated by the need to explain *causal inference* techniques that derive causal conclusions from patterns of unconditional and conditional correlations.

These causal inference techniques require us to suppose that:

> X causes Y iff it is an ancestor of Y in a recursive structure of (pseudo-)deterministic laws in which the exogenous terms are probabilistically independent of each other.

Illustration:
$X \leftarrow e_x$
$Y \leftarrow aX + e_y$
$Z \leftarrow bX + cY + e_z$

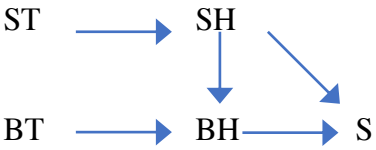
If we reordered these equations/laws so that, say, X depended on Y rather than vice versa, then the exogenous e_vs would no longer be probabilistically independent. (Intuitively, the idea is that any variable that causally depends on others will always have sources of variation that are probabilistically independent of those other variables. That's what shows it to be an effect of those other variables.)

We can generalise to any variables $X1, \ldots Xn$, and exogenous terms (r.h.s. only), $e1, \ldots en$ and any recursive laws (variables might be dichotomous, or determinable, or ordinal, as well as real-valued, and the laws not just linear):
$Xi \leftarrow F(X1, \ldots Xi\text{-}1, ei)$

References: "The Causal Structure of Reality" 2021 *PhilSci-Archive* 1-51
"The Statistical Nature of Causation" 2022 *The Monist* 105:247–75.

B Actual Causation and Counterfactuals

These recursive structures of laws with exogenous independence (RLIs) are by their nature generic. How do they relate to *single-case* causation and counterfactual dependence? Here I appeal to the wealth of work in the Lewisian tradition that has devised recipes for reading answers to these questions off from "causal models" that specify how dependent variables are deterministic functions of others.



What do the arrows mean here? Existing work takes them to portray either (a) some primitive causal dependence or (b) complex counterfactuals. I can do better: the models show how the variables are related in an RLI.

This opens the way to explaining counterfactuals (and actual causation) in terms of RLIs.

Interventionist semantics for counterfactuals: set the relevant variable to the counterfactual value, leave all non-descendants unchanged, crank the equations . . .

C Rational Action

When is it rational to do A in pursuit of B? I say: when it is probable that B "depends causal-counterfactually" on A. (Note how this explains rational action in terms of counterfactuals in terms of causation in terms of RLIs . . . I'm

going to say "causal-counterfactually" to remind us that I'm analysing counterfactuals in terms of a prior analysis of causation.)

The "probable" here is practically necessary. *Smoking causes cancer* and so it's rational not to smoke. But *your* avoiding cancer mightn't causal-counterfactually depend on *your* not smoking, and you'll never know enough to know for sure whether it does or not. The causal difference smoking makes to cancer is the *probability*, given your knowledge of your situation, that your cancer-avoidance does indeed causal-counterfactually depend on your non-smoking.

But the "depends causal-counterfactually" is also necessary. It's not enough that $Pr(B|A) > Pr(B)$. Car accidents are fewer among those who buy good car insurance, but that doesn't make it rational to buy good insurance to avoid accidents—because a mere correlation *doesn't* evidence any single-case causal-counterfactual dependence of accident-avoidance on insurance (as is shown by the way the correlation disappears when we control for the factor on which accident-avoidance *does* causal-counterfactually depend, viz. a cautious disposition).

D Agency Theories

Agency theorists will say this is all back-to-front. I am explaining rational action in terms of an independent counterfactual metaphysical dependency, the asymmetry of which is grounded in the asymmetry of RLIs. Wouldn't we do better to explain causation and its asymmetry in terms of when it's rational to act?

After all, my account doesn't do anything to explain *why* rational action is rational. What's so good about doing A in pursuit of B just in case . . . [some complex story involving RLIs]? Shouldn't we rather start with the simple idea that it's rational to do A in pursuit of B just in case doing A renders B likely, and then—hopefully—explain causation in terms of that?

Newcombe's paradox drives the point home. What's so good about acting on causes? If you're so smart, why aintcha rich? But of course Newcombe's paradox is a double-edged sword for agency theorists. Buying good car insurance renders accident-avoidance likely . . .

The next two sections will look at the options open to agency theorists and argue they don't work. I'll then turn to Albert and Loewer, who at first pass might seem to have much in common with agency theorists. I'll show, however, that they are much closer to my line—and, having shown this, I'll argue that my line is preferable to theirs.

E Freedom

Some (Glymour? Hitchcock? Woodward? Price?) pursue a bad line of thought: rational agents choose *freely*; so *their* decisions won't be correlated with the other causes (cautious disposition) of desired outcomes (few accidents); so they can conclude that their action A won't render B likely; so they can avoid acting on the spurious A-B correlation.

Hm. We can debate about exactly what free action requires . . . But it surely *won't* require that my cautious disposition doesn't make me more likely to buy expensive insurance than less prudent people. (Of course, in making a decision, we can *compute* the A-B influence *on the pretence* that A is *decorrelated from the other causes* of B—ie on the supposition that we're sort of super-free. But that's not because we really are super-free, but just because this is a good way to work out how likely it is that B is *counterfactually* dependent on A (which will be zero in spurious cases like insurance-accidents, since, as I observed earlier, good insurance adds no cases to those in which accident-avoidance already causal-counterfactually depends on caution).)

F Tickles

So good agency theorists take a different line—the tickle defence. They start from the good point that decision-theoretic reasoning should always work within the reference class defined by everything you *know* about your situation (cf principle of total evidence). And then they argue that you will always know something (eg that you're

cautious) that will render any non-effective A probabilistically irrelevant to B (among cautious people expensive insurance isn't associated with fewer accidents).

I don't think the tickle line works. Apart from placing unreasonable demands on conscious introspection, it also struggles with incontinent agents, whose actions are partly influenced by their decisions but also partly influenced by *subconscious* factors spuriously associated with the effect (cf Lewis). In such cases, there'll still be an A-B correlation in the reference class fixed by agents' knowledge of their motives, but it will clearly be irrational for them to be moved by this in their decision-making.

Tickle defenders typically respond by saying they want to focus on agents that are fully knowledgeable and continent. Maybe that is reasonable if they are just aiming to analyse the *concept* of causation (but isn't *continent* a causal notion? and anyway doesn't the concept of causation have many different psychological sources?). But I don't think this will do if we are after a relation that's self-evidently good to act on—after all, *correlated in the reference class defined by the agent's knowledge* just isn't good for incontinent agents to act on.

So I think we should explain rational action in terms of causation, rather than vice versa. It's not rational to buy expensive insurance to prevent accidents (when you know the correlation is spurious) simply because the spuriousness of the correlation tells you that accident-avoidance is never causal-counterfactually dependent on buying insurance. And you can know this without attending to tickles and placing yourself in some narrow reference class.

G Albert and Loewer

David Albert and Barry Loewer have views which at first sight look like the agency line. But in truth they are much closer to me.

They don't do counterfactuals in terms of causation, as I do, but go straight to counterfactuals (and aim to explain causation in terms of that).

They assume a metaphysics of deterministic statistical mechanics which they call "The Mentaculus". The universe starts in a low entropy macrostate ("the past hypothesis") and then evolves into macrostates of increasing entropy.

Albert and Loewer say (very roughly) that:

> B counterfactually depends on A iff removing A from actuality but keeping current macro-facts fixed will, given basic dynamics and the past hypothesis, imply the absence of B.

They hold that many forward-facing counterfactuals (B after A) are true, but no (or few) backward-facing ones. This asymmetry derives from the asymmetry of (near) overdetermination: the present contains many macro-traces of the past, but none of the future (cf Lewis). So having not-A rather than A, plus present macro-facts, can make a big difference to the probability of future events but not (normally) to past events.

Note how this line will also allow them to explain why it's not rational to act on spurious correlations like that between buying good insurance and accident-avoidance. Since the earlier cautious disposition will have left many present macro-traces, the insurance won't make any probabilistic difference in the reference class defined by those traces. Just as macro-records screen off past events from present actions, so do they screen off future events that are spuriously correlated with present actions.

This might make it look like Albert and Loewer are akin to agency theorists, getting counterfactuals (and causation) simply by considering which antecedents/actions make results B likely in the relevant reference classes.

But not so. The crucial point here is that their reference classes aren't defined by *agents' knowledge*, but by *all present macro-facts*. Many of these macro-facts might well be unknown to agents. (Note that Albert and Loewer never argue that we'll *know* all current macro-facts—they don't need to, since they are happy to screen off effects from spurious antecedents by *current* macro-facts, not by *known* macro-facts.)

So Albert and Loewer come out like me here. Rational agency is explained in terms of actions making a counterfactual difference, and this is explained in terms of objective metaphysical patterns. Spurious action-result correlations that remain given agents' knowledge are no problem for them, nor for me, since they never make a counterfactual difference.

H Albert and Loewer vs Papineau

The difference between Albert and Loewer and me thus lies in how we do counterfactuals. I do it in terms of asymmetric RLIs and the resulting recipe for reading off causal-counterfactuals. They do it more simply in terms of differences implied by given current macro-facts.

Albert and Loewer might seem to have an advantage over me here, resulting from the similarity they do bear to agency theories. Go back to the complaint that my account doesn't do anything to explain *why* rational action is rational. As I put the challenge earlier, what's so good about doing A in pursuit of B just in case . . . [some complex story involving RLIs]? Albert and Loewer look better off on this score. Isn't there something intuitive about the idea that you should perform actions that make desired results likely in the reference class *defined by present macro-facts*?

Well, maybe. But if that's so, I can appeal to the intuition too. This is because the asymmetry of (near-) overdetermination is a corollary of my own RLI account of causation. The various effects of a given cause will inevitably be correlated, given the independence of the exogenous terms. Which is to say that the many different effects will tend to co-occur whenever the cause does. So I too can say that it's a good idea to act on causes because then you'll generally be making desired results likely given present macro-facts.

I No-Trace Cases

Still, is it self-evident that it's a good idea to perform actions that make desired results likely in the reference class *defined by present macro-facts*? Where did that come from? Why specifically *present* and *macro-facts*? Let me finish by putting pressure on this idea.

As is familiar, Albert and Loewer face difficulties in connection with those (overwhelmingly unlikely but naturally possible) cases where past events leave no present traces. They are forced to bite the bullet and say that in such cases there can be counterfactual dependence of past macro-facts on present events, and they seek to lessen the counter-intuitiveness by observing that in such cases agents will be in no position to exploit this in action.

Note than on my account such cases don't yield causal-counterfactual dependence. Even if we freakily have a case where all the common effects of some cause all fail to occur, my asymmetric lawlike structures in the form of RLIs will still apply and imply causal-counterfactual dependencies accordingly. So, on my account, Atlantis's existence won't causal-counterfactually depend on my not clicking my fingers, nor my accident-avoidance on my buying insurance, even if Atlantis or my caution leave no traces to screen off these results from these actions.

So perhaps we shouldn't have been so quick to accept the idea we should perform actions that make desired results likely in the reference class *defined by present macro-facts*. The issue is clearest with forward-facing spurious correlations. Focus on the case where my caution happens to leave no present traces. Albert and Loewer will be forced to say my accident-avoidance now counterfactually depends on my insurance-buying, and will seek somehow to explain this away (presumably by reference to the unknowability of such unlikely counterfactual dependencies). But why not join me and simply say there's no causal-counterfactual dependence, because of the structure of the underlying RLI laws?

So the truth, I'd say, is *not* that we should perform actions which make desired results likely given present macro-facts (and still less those which make them likely given agents' knowledge), but rather perform actions on which desired results probably *counterfactually depend* in my causal sense.